Matching and Inference for Multiple Correlated Data Sets

by

Cencheng Shen

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2015

Abstract

Given multiple correlated data sets, an important question is how to make use of them to benefit later statistical inference. This is a realistic setting in the modern world as more and more related data sets are collected, say images and their descriptions, articles in multiple languages, actors in multiple social networks; and real data are often multivariate or high-dimensional such that dimension reduction is necessary before any inference.

In this dissertation, I consider three dimension reduction and matching methods, namely principal component analysis followed by Procrustes matching, canonical correlation analysis, and nonlinear matching using shortest-path distance and joint neighborhood. I investigate their theoretical properties and their impact on later inference using the Procrustes fitting error, classification error, and hypothesis testing respectively.

The main conclusion of this dissertation is that given a particular inference task for multiple correlated data sets, we may significantly improve the inference performance by joint matching and projection, compared to separate projection or omitting

ABSTRACT

modalities. Numerical experiments are provided to illustrate the theorems and the methodology using simulated data and real data.

Primary Reader: Dr. Carey E. Priebe

Secondary Reader: Dr. Minh Tang

Acknowledgments

First, I would like to thank my advisor Professor Carey Priebe. He has been very supportive throughout my graduate life, and offered many insightful suggestions in improving this dissertation and related papers. I am extremely grateful for this opportunity to know and work with him; all the invaluable lessons within and beyond academics from him make me a better person in learning, research, communications, and characteristics.

Second, I would like to thank all the persons that collaborated with me in relevant papers and helped me in this dissertation: I spent the first two years working with Dr. Donniell Fishkind on my very first research project – the incommensurability phenomenon, which is a very rewarding experience for my thinking and writing in the long run; Dr. Ming Sun worked together with me on the canonical correlation project; Dr. Minh Tang has been very helpful in various aspects of my research, as well as being a nice and fun person to talk to; Dr. Youngser Park provided many technical and numerical supports.

Thirdly, I would like to thank many other persons involved in my research and

career path: Dr. Joshua Vogelstein and Li Chen have worked with me on other projects, which made my research experience more colorful and diversified. Dr. Daniel Sussman, Dr. Sancar Adali, Dr. Vince Lyzinski, etc., all discussed and shared with me various research topics and interesting directions. There are many memorable teachers and students I met in China, Singapore, the U.S. during my high-school, undergraduate, and graduate study, and I thank them for their guidance, friendship, and encouragement along my path.

Last but most, my sincere thanks to my parents, because they prepare me better in this wonderful yet difficult world.

Dedication

This thesis is dedicated to my mother, without whom I would not be who I am.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the modern world it is becoming increasingly important to deal effectively with large amounts of high-dimensional data. For the purpose of data analysis, it is imperative to consider dimension reduction and embed the data into a low-dimensional space for subsequent analysis. Traditional linear embedding techniques have solid theoretical foundations and are widely used, e.g., principal component analysis (PCA) [1], [2] and multi-dimensional scaling (MDS) [3], [4], [5] for a single data set, and canonical correlation analysis (CCA) [6], [7] for multiple data sets.

However, real data may exhibit nonlinear geometry, and unfolding the non-linearity can be beneficial for subsequent inference. Recently many manifold learning algorithms have been proposed to learn the intrinsic low-dimensional structure of nonlinear data, including Isomap [8], [9], locally linear embedding (LLE) [10], [11], Hessian LLE [12], Laplacian eigenmaps [13], [14], local tangent space alignment (LTSA) [15],

[16], among many others. Most algorithms start with the assumption that the data are locally linear, and explore the local geometry via the nearest-neighbor graph of the sample data: transformation of the data is carried out based on the neighborhood graph, and the low-dimensional manifold is learned by optimizing some objective function. These nonlinear embedding algorithms usually serve as a preliminary feature extraction step enabling subsequent inference, and have achieved many practical successes in object recognition, image processing, etc.

Despite the number of available data processing techniques, their impact for later inference is usually less clear. For example, when given multiple correlated data sets, an important question regarding robustness is: if the original data sets are highly correlated, will the projected data sets still be highly correlated? And is there any information loss after dimension reduction? We tackle this question in Chapter 2 by investigating the Procrustes fitting error of two correlated data sets after separate PCA projection. In turns out that depending on the covariance structure, the Procrustes error may be larger than usual and cause information loss after separate projection. We name it as the incommensurability phenomenon, and the content of Chapter 2 is based on our paper [17].

The incommensurability phenomenon can be avoided if joint projection (say CCA) is used rather than separate PCA projection. A natural follow-on question will be: can we quantify the inference advantage of joint projection over separate projection? Or alternatively, can we always improve the later inference by collecting and utilizing

more data sets? We consider the classification task in Chapter 3, and use generalized canonical correlation analysis (GCCA) to jointly project the data sets. Indeed the classification performance may be significantly improved by using more data sets, assuming the extra data sets satisfy certain similarity condition. This chapter is based on our paper [18].

Then in Chapter 4 we consider how to incorporate nonlinear embedding algorithms into our matching and projection framework. We present a nonlinear manifold matching algorithm to match multiple data sets using shortest-path distance and joint neighborhood selection. This is effectively achieved by combining Isomap [8] and the matching methods from [19]. Our approach exhibits superior and robust performance for matching data from disparate sources, compared to algorithms that do not use shortest-path distance or joint neighborhood selection; in particular, we use distance correlation [20] and hypothesis matching test as our evaluation criteria. This chapter is based on our paper [21].

Chapter 5 briefly concludes this dissertation.

Chapter 2

The Incommensurability Phenomenon for Separate Projections

2.1 Introduction

In this chapter we investigate the Procrustes fitting error between two matrices $\mathcal{X}$ and $\mathcal{Y}$:

$$\epsilon = \min_{Q'Q=I} \|Q\mathcal{X} - \mathcal{Y}\|_F, \tag{2.1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ represent low-dimensional projections of two multivariate data sets, Q is a rotation matrix and $'$ is the transpose sign. We show that the square Procrustes fitting error is asymptotically a convex combination (via a correlation parameter) of

the Hausdorff distance between the projection subspaces and the maximum possible value of the square Procrustes fitting error for normalized data. It turns out that two separately projected data sets may have an inordinately large Procrustes fitting error, even when the original data sets are closely correlated with each other, which is called as the incommensurability phenomenon.

We start with some background information in Section 2.2, followed by an idealized example in Section 2.3 so as to introduce the relationship between the Procrustes fitting error ϵ and the Hausdorff distance in a simplified setting. In Section 2.4 we consider their relationship under more general scenarios, and quantify the incommensurability phenomenon. Numerical experiments are presented in Section 2.5. Discussions follow in Section 2.6, and proofs are in Section 2.7.

Note that this chapter is based on the paper [17].

2.2 Background

As already mentioned in Chapter 1, dimension reduction is often applied to modern data before any later inference; thus it is very practical to compare the projected data rather than the original data. And we confine ourselves to principal components analysis (PCA), which remains a very popular and successful method to process the data.

For the comparison of two data sets, the Procrustes fitting error is a simple yet

useful statistic to use. To give just a few examples, see [22] and [23] where Procrustes fit is used to assess the goodness-of-fit between two slightly different spatial configurations projected to a lower dimensional space by multi-dimensional scaling. Procrustes analysis is similarly seen to be a valuable tool in [24], [25], [25], [26], [19], [27], and is also used for matching purpose in Chapter 4.

Therefore, this chapter is devoted to separate PCA projections of two correlated multi-dimensional data followed by Procrustes analysis; and the incommensurability phenomenon may occur depending on the covariance structure.

2.3 A Cautionary Tale of Two Scientists

In this section we explore an idealized scenario for the purpose of straightforward illustration; the general setting will be treated in Section 2.4.

Suppose that two scientists each take measurements of m features of certain random process, where m is a large, positive integer. For each $i = 1, 2, 3, \ldots$ (such as i days or i objects), the first scientist records her measurements as $\mathbf{X}_i \in \mathcal{R}^m$ ($\mathbf{X}_{ij}$ is the measurement on jth feature for $j = 1, 2, \ldots, m$); and the second scientist records his measurements as $\mathbf{Y}_i \in \mathcal{R}^m$, etc. The two scientists should be measuring the same or at least two similar processes, but $\mathbf{X}_i$ is usually not the same as $\mathbf{Y}_i$.

For each positive integer n, denote by $X^{(n)}$ the data matrix $[\mathbf{X}_1|\mathbf{X}_2|\cdots\mathbf{X}_n] \in \mathcal{R}^{m\times n}$ consisting of the first scientist's measurements over n (days or objects), and

denote by $Y^{(n)}$ the data matrix $[\mathbf{Y}_1|\mathbf{Y}_2|\cdots\mathbf{Y}_n] \in \mathcal{R}^{m\times n}$ consisting of the second scientist's measurements. Let us assign a notational distribution for each collection of measurements such that (s.t.) $\mathbf{X}_i \overset{i.i.d.}{\sim} X$ and $\mathbf{Y}_i \overset{i.i.d.}{\sim} Y$; doing so implies each collection of data $\{\mathbf{X}_i\}$, $\{\mathbf{Y}_i\}$ are independently identically distributed (i.i.d.) within each collection.

We denote the covariance matrix of $[X, Y]$ as

$$\Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \\ \Sigma'_{XY} & \Sigma_Y \end{bmatrix} \in \mathbb{R}^{2m\times 2m}.$$

And in this simple tale let us consider an ideal case of the covariance matrix: $\Sigma_X = \Sigma_Y = I$ and $\Sigma_{YX} = \rho I$. So ρ is the correlation. If $\rho = 1$, the two scientists' measurements are exactly the same for all i and $\mathbf{X}_i = \mathbf{Y}_i$; and if $\rho = 0$, $\mathbf{X}_i$ are uncorrelated with $\mathbf{Y}_i$. We will generalize the covariance matrix and ρ in Section 2.4.

As the data dimension is usually high, we assume that the scientists project the data to a lower-dimensional space $\mathcal{R}^d$ (d is a preset positive integer no larger than m). This is done as follows: Let $H_n = I_n - J_n/n$ denote the centering matrix (I_n and J_n are, respectively, the $n \times n$ identity matrix and the matrix of all ones), and denote $\mathcal{G}_{d,m}$ as the Grassmann manifold (i.e., the space of all d-dimensional subspaces of $\mathcal{R}^m$, or called d-planes for simplicity). Suppose for each n, the first scientist chooses a subspaces $\mathcal{A}^{(n)}$ in $\mathcal{G}_{d,m}$ and the second scientist chooses $\mathcal{B}^{(n)}$ in $\mathcal{G}_{d,m}$. The choice of such d-planes forms two sequences $\left\{\mathcal{A}^{(n)}\right\}$ and $\left\{\mathcal{B}^{(n)}\right\}$. In this section we do not

specify how the projection subspaces are chosen for each n, but in the remaining sections they are selected by PCA.

Note that for any d-plane $\mathcal{A}$ in $\mathcal{G}_{d,m}$, there corresponds a unique $m \times m$ orthogonal projection matrix $P_{\mathcal{A}}$ idempotent of rank d, and the projected low-dimensional data is $P_{\mathcal{A}^{(n)}} X^{(n)} H_n$ and $P_{\mathcal{B}^{(n)}} Y^{(n)} H_n$. By using the size m idempotent projection matrices, we opt to keep the ambient coordinate system of $\mathcal{R}^m$ for the projected data's range instead of $\mathcal{R}^d$; this practice does not affect ϵ and any other result in this chapter, and facilitates later statements and proofs.

Thus at any given n, the two centered, projected, scaled data sets reported to the Governing Board of Scientists by the first and the second scientists are:

$$\mathcal{X} = \frac{\sqrt{d}}{\|P_{\mathcal{A}^{(n)}} X^{(n)} H_n\|_F} P_{\mathcal{A}^{(n)}} X^{(n)} H_n \in \mathbb{R}^{m \times n}$$
$$\mathcal{Y} = \frac{\sqrt{d}}{\|P_{\mathcal{B}^{(n)}} Y^{(n)} H_n\|_F} P_{\mathcal{B}^{(n)}} Y^{(n)} H_n \in \mathbb{R}^{m \times n}.$$

Now the Governing Board of Scientists wants to perform its own check that the two scientists are indeed taking similar measurements, by testing the Procrustes fitting error ϵ between $\mathcal{X}$ and $\mathcal{Y}$. It is not hard to show that $0 \leq \epsilon \leq \sqrt{2d}$, and they decide to accept the null assumption that the two collections of data are matched with each other, if and only if the statistic ϵ is small (for example, test against the alternative that the two collections of data are pairwise independent). Is this a proper test?

Before answering, let us define the Hausdorff distance between any two d-planes

$\mathcal{A}$ and $\mathcal{B}$ as

$$d(\mathcal{A},\mathcal{B}) = \sqrt{\sum_{i=1}^{d}(2\sin(\frac{\theta_i(\mathcal{A},\mathcal{B})}{2}))^2} = \sqrt{\sum_{i=1}^{d}2(1-\cos\theta_i(\mathcal{A},\mathcal{B}))}, \tag{2.2}$$

where $\theta_i(\mathcal{A},\mathcal{B})$ denotes the ith increasingly ordered principal angle between the two PCA subspaces. The Hausdorff metric is unitarily invariant (see to [28]), and takes value in $[0,\sqrt{2d}]$. Note that even though the Hausdorff distance notation is similar to the projection dimension, there should not be any confusion of them in this chapter.

Then we have the following theorem about the Procrustes fitting error and the Hausdorff distance under this simplified setting:

Theorem 1. *For any two sequences of arbitrary projection subspaces* $\left\{\mathcal{A}^{(n)}\right\}$ *and* $\left\{\mathcal{B}^{(n)}\right\}$,

$$\epsilon^2 - [(1-\rho)\cdot 2d + \rho\cdot d^2(\mathcal{A}^{(n)},\mathcal{B}^{(n)})] \overset{a.s.}{\to} 0. \tag{2.3}$$

The proof of Theorem 1 is in fact a special case of the more general Theorem 2. Theorem 1 says that ϵ^2 is asymptotically the convex combination of $2d$ and $d^2(\mathcal{A}^{(n)},\mathcal{B}^{(n)})$ via ρ. In particular, if the two scientists' measurements are independent from each other, ϵ approaches the maximum because $\rho = 0$; if the scientists' measurements are almost the same, ρ is close to 1 and ϵ is close to $d(\mathcal{A}^{(n)},\mathcal{B}^{(n)})$, which means that ϵ may be large depending on the choice of $\mathcal{A}^{(n)}$ and $\mathcal{B}^{(n)}$. Is the Hausdorff distance close to zero when ρ is close to 1?

Indeed in Section 2.4 we show that when separate PCA projections are used, the Hausdorff distance and thus the Procrustes fitting error may even be close to

the maximum value $\sqrt{2d}$. This unfavorable behavior is shown in Section 2.5, which degrades the power of a matching test and potentially hampers the decision of the Governing Board.

In contrast, in this simple tale, if the two scientists decide a common projection such that $\mathcal{A}^{(n)} = \mathcal{B}^{(n)}$, then $d(\mathcal{A}_n, \mathcal{B}_n) = 0$. Thus a higher ρ always implies a smaller value of ϵ asymptotically. But this is not necessarily a favorable strategy, see in Section 2.6.

2.4 Main Results

In this section, we keep all previous setting the same, but we allow the covariance matrix of $[X, Y]$ to be arbitrary. Furthermore, from now on $\mathcal{A}^{(n)}$ and $\mathcal{B}^{(n)}$ are the respective d-planes to which PCA projects the centered data matrices $X^{(n)}H_n$ and $Y^{(n)}H_n$.

Next we define a weighted form of the Hausdorff distance

$$\tilde{\eth}(\mathcal{A}^{(n)}, \mathcal{B}^{(n)}) = \sqrt{\sum_{i=1}^{d} 2\left(1 - \frac{1}{\frac{1}{d}\sum_{j=1}^{d}\sigma_j\left(\Sigma_{XY}\right)}\sigma_i\left(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}\right)\right)}, \tag{2.4}$$

where $\sigma_i(C)$ denotes the ith decreasingly ordered singular values of any matrix C.

Note that if we consider the setting in Section 2.3, $\tilde{\eth}^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})$ is equivalent to $d^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})$, because $\sigma_i(P_{\mathcal{A}}P_{\mathcal{B}}) = \cos\theta_i(\mathcal{A}, \mathcal{B})$ for any two subspaces $\mathcal{A}$ and $\mathcal{B}$; and the weighted Hausdorff distance also takes value in $[0, \sqrt{2d}]$ as shown by Proposition 5.

Now we present a general theorem regarding the Procrustes fitting error ϵ between

the separate PCA projected data $\mathcal{X}$ and $\mathcal{Y}$:

Theorem 2. *It holds almost surely that*

$$\epsilon^2 - \left[(1-\rho)\cdot 2d + \rho \cdot \eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})\right] \to 0 \tag{2.5}$$

as $n \to \infty$, where ρ is defined as

$$\rho = \frac{\sum_{j=1}^{d} \sigma_j(\Sigma_{XY})}{\sqrt{\sum_{j=1}^{d} \sigma_j(\Sigma_X)}\sqrt{\sum_{j=1}^{d} \sigma_j(\Sigma_Y)}}.$$

Note that the ρ here is a generalization of the same notation in Section 2.3, and we will show that $0 \le \rho \le 1$ in Proposition 6.

For small ρ, ϵ is close to $\sqrt{2d}$; and for large ρ, ϵ is close to $\sqrt{\eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})}$. Also if the Hausdorff distance is close to 0, ϵ is close to a fixed constant $\sqrt{(1-\rho)\cdot 2d}$. Therefore, as long as the Hausdorff distance is small, the Procrustes error will be close to a constant; but if it is large, the Procrustes error may be close to its maximum, giving rise to the incommensurability phenomenon.

A natural question is, when and how the Hausdorff distance $d(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})$ (or $\eth$) can be large? When there are repeating eigenvalues in Σ_X, and the dimension choice d cuts into the repeated eigenvalue, then the PCA subspace $\mathcal{A}^{(n)}$ will be uniformly distributed in the eigenspace corresponding to the repeated eigenvalue [29], [30]; the same holds for $\mathcal{B}^{(n)}$. Due to this randomness of PCA projection, we may have large Hausdorff distance even if the original data are highly correlated. The simulation in Section 2.5 provides an example.

This phenomenon makes two similar data sets look less similar after separate projection, though the exact impact is not easy to quantify especially for real data. But we do provide a real data example in [17]; and there are many ways to avoid the incommensurability phenomenon, which we will discuss in Section 2.6.

2.5 Numerical Experiments

We numerically validate the asymptotic relationship between the Procrustes fitting error and the Hausdorff distance, and show the incommensurability phenomenon.

We consider the following matched setting $X^{(n)} = Z^{(n)} + E_1^{(n)}$, $Y^{(n)} = Z^{(n)} + E_2^{(n)}$, where each object (row) of $Z^{(n)}$ is i.i.d. Gaussian distributed with $\Sigma_Z = \Sigma_{XY} = diag(9, 9, 4, c, c, c, 1, 1, 1)$, and each row of $E_i^{(n)}$ is just white noise with identity variance. Thus $\Sigma_X = \Sigma_Y = diag(10, 10, 5, c+1, c+1, c+1, 2, 2, 2)$, and $[X, Y]$ follows a multivariate normal distribution. We consider two diagonal variance settings with c set to be 1 and 4 respectively.

We generate the data for $n = 400$, choose $d = 3$, project the $X^{(n)}$ and $Y^{(n)}$ by PCA, and compute the Procrustes fitting error. It is carried out for 1000 Monte-Carlo replicates.

In Figure 2.1 and Figure 2.2 we present the numerical behaviors. The left part of both figures is to show the convergence behavior of ϵ as claimed in Theorem 2: the Hausdorff distance is the x-axis, based on which we draw a green line predicting ϵ by

Equation 2.5 (called Hausdorff distance equation in the figure), and the y-axis is the Procrustes error; then for each Monte-Carlo run, we calculate the Hausdorff distance and the Procrustes error of that run, draw a red point in the figure. Clearly the red points are very close to the green line, implying the correctness of our theorem. To better illustrates the convergence behavior, Table 1 also lists the maximum absolute difference between ϵ and the Hausdorff distance equation for both scenarios at $n = 50, 100, 200, 400$ (the convergence rate is approximately square root of n).

The right part of both figures is the histogram of the Hausdorff distance. Clearly in case of $c = 1$, the PCA subspaces are very close to each other (both converge to the subspace spanned by the first d coordinate axes), so that the Hausdorff distance is close to zero and the Procrustes error is almost a constant. But in case of $c = 4$, the repeated eigenvalue of Σ_X and Σ_Y at $d = 3$ causes certain degree of randomness for the two PCA subspaces, so that the Hausdorff distance and the Procrustes error are larger than necessary, which reflects the incommensurability phenomenon.

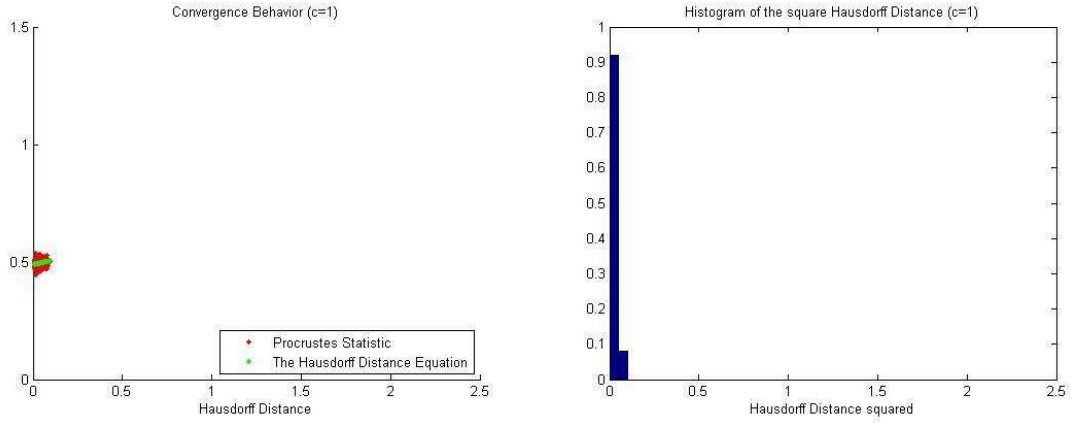


Figure 2.1: The Diagonal Variance (c=1) Setting

Table 2.1: The maximum absolute difference between ϵ and the Hausdorff distance equation over 1000 iterations

	n=50	n=100	n=200	n=400
Diagonal Variance (c=1)	0.1479	0.0984	0.0666	0.0461
Diagonal Variance (c=4)	0.1485	0.1063	0.0731	0.0518

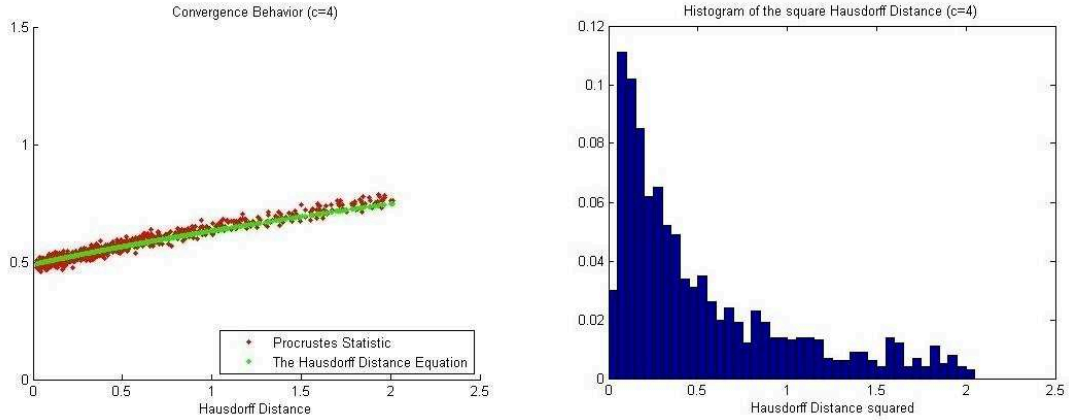


Figure 2.2: The Diagonal Variance (c=4) Setting

Next we check the effect of the incommensurability phenomenon for hypothesis testing, which is the same test in the matching framework of [19] or Chapter 4.

We use the same matched distribution of the previous example, and first generate $n = 200$ matched training pairs to learn the projections and the matching, then generate 200 matched testing pairs to obtain the empirical distribution of the testing statistic under the null. It is tested against the same number of generated unmatched testing pairs to give the test power, and for the unmatched pairs each column of $X^{(n)}$

Table 2.2: The Matching Test Powers for Diagonal Variance Setting

	Omnibus	CCA	P ∘ M
$c = 1$ with $\alpha = 0.05$	0.8033	0.7621	0.8015
$c = 4$ with $\alpha = 0.05$	0.7954	0.7689	**0.7130**
$c = 1$ with $\alpha = 0.15$	0.8846	0.8630	0.8838
$c = 4$ with $\alpha = 0.15$	0.8782	0.8663	**0.8271**

and $Y^{(n)}$ are i.i.d. jointly multivariate normal with the same variances but covariance 0 (namely Σ_X and Σ_Y are the same as in the null but $\Sigma_{YX} = 0$). The average powers of 1000 Monte-Carlo replicates are plotted in Figure 2.3, for P ∘ M, joint MDS (also called Omnibus) and CCA matching methods. (see in [19] or Chapter 4 for more details of the matching methods.)

The constant c in the variance is again used to control whether incommensurability phenomenon happens or not: if $c = 1$ the incommensurability phenomenon is avoided, while it happens at $c = 4$. So the power degradation of P ∘ M at $c = 4$ is indeed expected in Figure 2.3, compared to the $c = 1$ case and the other two joint matching methods that are always immune to the incommensurability phenomenon. The effect is also clear from Table 2, for which we list the matching powers at critical value $\alpha = 0.05$ and $\alpha = 0.15$ for all methods. This experiment indicates that the incommensurability phenomenon affects hypothesis testing.

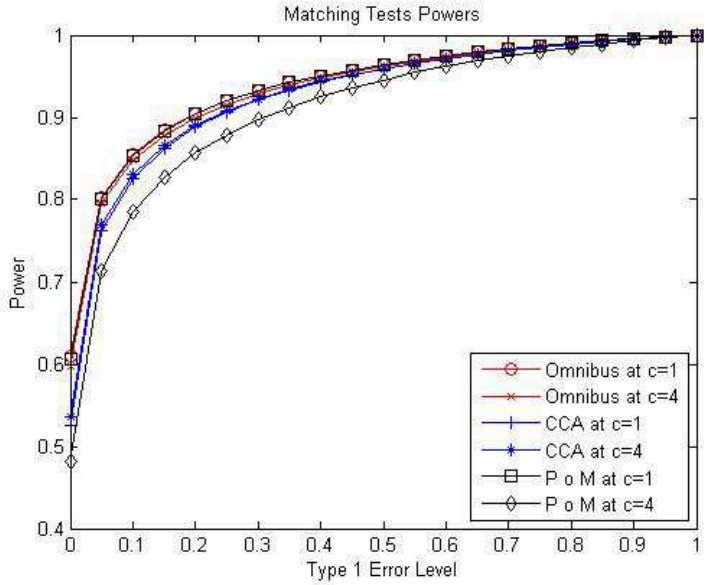


Figure 2.3: Matching Tests for Diagonal Variance Settings

2.6 Discussions

Several factors may contribute to the manifestation of the incommensurability phenomenon when two correlated data sets are projected to a lower dimension. One factor is the circumstance where the two data sets are projected separately when the dimension reduction is performed. Another factor is the circumstance where the choice of embedding dimension d does not provide for a sufficiently large gap between the dth and $d+1$th eigenvalues of the covariance matrix for the data sets. These factors may combine to allow substantial probability of having significant distance between the separate projection subspaces, which then causes an inordinately large Procrustes fitting error.

Of course, as we have seen in the numerical experiment and the proof, one remedy

is simply not to do the projections separately for the two data sets; robust joint embedding schemes are available, such as developed in [31], [32], and [19]. Indeed, an easily used candidate is canonical correlation analysis (CCA) [6], [33], which can be extended to situations where more than two data sets are being treated, and CCA has good properties for subsequent inferential tasks [34], [35], [18], and our Chapter 3. The incommensurability phenomenon can then be avoided at the cost of the extra computation involved, although this extra computation may be a significant burden when dealing with a large volume of data in a distributed system.

Another possible remedy is to choose the embedding dimension d so as to maintain enough of a gap between the dth and $d+1$th eigenvalues of the data sets' covariance matrix, or simply set the two projections to be the same. However, this remedy is not always useful, limits the embedding dimension, and may come at the expense of abandoning additional signal just for the matching purpose.

There are many interesting future directions to consider from this chapter. For example, it may be worthwhile to find the distribution of the Hausdorff distance between PCA projections of correlated data; also a high-dimensional case can be very useful in practice, for which the incommensurability phenomenon should be more serious; and we would like to extend the incommensurability phenomenon to other dimension reduction methods other than PCA, say sparse PCA or even nonlinear projections.

2.7 Proofs

We can expand Equation 2.1 into

$$\epsilon^2 = \|\mathcal{X}^{(n)}\|_F^2 + \|\mathcal{Y}^{(n)}\|_F^2 - 2\sum_{i=1}^{m}\sigma_i(\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime}) = 2d - 2\sum_{i=1}^{m}\sigma_i(\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime}). \tag{2.6}$$

In order to related ϵ to the weighted Hausdorff distance $\eth$ as in Theorem 2, we first establish Lemmas 3 and Lemma 4.

Lemma 3. *Almost surely,* $\text{trace}\frac{1}{n-1}P_{\mathcal{A}^{(n)}}X^{(n)}H_nH_n^{'}X^{(n)\prime}P_{\mathcal{A}^{(n)}}^{'} \to \sum_{i=1}^{d}\sigma_i(\Sigma_X)$ *as* $n \to \infty$.

Proof. For each $n = 1, 2, 3, \ldots$, let us consider the singular value decomposition

$$X^{(n)}H_n = U^{(n)}\Lambda^{(n)}V^{(n)\prime}$$

where $U^{(n)} \in \mathbb{R}^{m\times m}$ is orthogonal, $\Lambda^{(n)} \in \mathbb{R}^{m\times n}$ is a diagonal matrix with the singular values being non-increasing along its diagonal, and $V^{(n)} \in \mathbb{R}^{n\times n}$ is orthogonal.

By the definition of PCA,

$$P_{\mathcal{A}^{(n)}}X^{(n)}H_n = U^{(n)}E\Lambda^{(n)}V^{(n)\prime},$$

where $E \in \mathbb{R}^{m\times m}$ is the diagonal matrix with its first d diagonals being 1 and its remaining diagonals being 0. Thus, the matrix

$$X^{(n)}H_nH_n^{'}X^{(n)\prime} = U^{(n)}\Lambda^{(n)}\Lambda^{(n)\prime}U^{(n)\prime}$$

and the matrix

$$P_{\mathcal{A}^{(n)}}X^{(n)}H_nH_n^{'}X^{(n)\prime}P_{\mathcal{A}^{(n)}}^{'} = U^{(n)}E\Lambda^{(n)}\Lambda^{(n)\prime}E\ U^{(n)\prime}$$

share their d largest singular values, with the remaining $m-d$ singular values of the latter matrix being 0. By the strong law of large numbers, almost surely

$$\frac{1}{n-1}X^{(n)}H_nH_n'X^{(n)\prime} \to \Sigma_X,$$

hence we have

$$\text{trace}\frac{1}{n-1}P_{\mathcal{A}^{(n)}}X^{(n)}H_nH_n'X^{(n)\prime}P_{\mathcal{A}^{(n)}}' \to \sum_{i=1}^{d}\sigma_i(\Sigma_X)$$

as $n \to \infty$.

Lastly, recall that in Section 2.3 we explicitly allow $\{\mathcal{A}^{(n)}\}_{n=1}^{\infty}$ to be any elements of $\mathcal{G}_{d,m}$ in the special case that $\Sigma_X = \alpha \cdot I_m$ for some $\alpha > 0$; indeed, in this special case we also have

$$\begin{aligned}
&\text{trace}\frac{1}{n-1}P_{\mathcal{A}^{(n)}}X^{(n)}H_nH_n'X^{(n)\prime}P_{\mathcal{A}^{(n)}}' \\
&= \alpha \cdot \text{trace}P_{\mathcal{A}^{(n)}} + \text{trace}P_{\mathcal{A}^{(n)}}\left(\frac{1}{n-1}X^{(n)}H_nH_n'X^{(n)\prime} - \alpha \cdot I_m\right)P_{\mathcal{A}^{(n)}}' \\
&\to \alpha d = \sum_{i=1}^{d}\sigma_i(\Sigma_X),
\end{aligned}$$

as $n \to \infty$, by the boundedness of $\{P_{\mathcal{A}^{(n)}}\}_{n=1}^{\infty}$ and the strong law of large numbers. □

Lemma 4. *For $i = 1, 2, \ldots, m$, almost surely*

$$\sigma_i^2(\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime}) - \delta \cdot \sigma_i^2(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}) \to 0$$

as $n \to \infty$, where $\delta = \frac{1}{\frac{1}{d}\sum_{j=1}^{d}\sigma_j(\Sigma_X) \cdot \frac{1}{d}\sum_{j=1}^{d}\sigma_j(\Sigma_Y)}$.

Proof. For each $n = 1, 2, \ldots$, expanding the expression $\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime}(\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime})'$ by the definition, we can write it as $\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime}(\mathcal{Y}^{(n)}\mathcal{X}^{(n)\prime})' = \phi^{(n)} \cdot \Phi^{(n)}$ where $\phi^{(n)}$ and $\Phi^{(n)}$

are defined as

$$\phi^{(n)} = \frac{d^2}{\text{trace}\frac{1}{n-1}P_{\mathcal{B}^{(n)}}Y^{(n)}H_nH_n'Y^{(n)\prime}P_{\mathcal{B}^{(n)}}' \cdot \text{trace}\frac{1}{n-1}P_{\mathcal{A}^{(n)}}X^{(n)}H_nH_n'X^{(n)\prime}P_{\mathcal{A}^{(n)}}'}$$

and

$$\Phi^{(n)} = P_{\mathcal{B}^{(n)}}\left(\frac{1}{n-1}Y^{(n)}H_nH_n'X^{(n)\prime}\right)P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\left(\frac{1}{n-1}X^{(n)}H_nH_n'Y^{(n)\prime}\right)P_{\mathcal{B}^{(n)}}'.$$

Let us also define $\Psi_{X,Y}^{(n)} = \frac{1}{n-1}X^{(n)}H_nH_n'Y^{(n)\prime} - \Sigma_{XY}$, for which we have $\Psi_{X,Y}^{(n)} \to 0$ as $n \to \infty$ by the strong law of large numbers. Thus, by the sub-additivity and sub-multiplicativity of the norm, and by the boundedness of $\{P_{\mathcal{A}^{(n)}}\}_{n=1}^{\infty}$ and $\{P_{\mathcal{B}^{(n)}}\}_{n=1}^{\infty}$, we have almost surely that

$$\begin{aligned}\|\Phi^{(n)} - P_{\mathcal{B}^{(n)}}\Sigma_{XY}'P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}'\|_F = \|&P_{\mathcal{B}^{(n)}}\Psi_{X,Y}^{(n)\prime}P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Psi_{X,Y}^{(n)}P_{\mathcal{B}^{(n)}}' \\ &+ P_{\mathcal{B}^{(n)}}\Psi_{X,Y}^{(n)\prime}P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}' \\ &+ P_{\mathcal{B}^{(n)}}\Sigma_{XY}'P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Psi_{X,Y}^{(n)}P_{\mathcal{B}^{(n)}}'\|_F \to 0,\end{aligned}$$

as $n \to \infty$.

Now, by Lemma 3 and the definition of $\phi^{(n)}$, almost surely $\phi^{(n)} \to \delta$ as $n \to \infty$, hence by the boundedness of $\{P_{\mathcal{A}^{(n)}}\}_{n=1}^{\infty}$ and $\{P_{\mathcal{B}^{(n)}}\}_{n=1}^{\infty}$, we have almost surely that

$$\begin{aligned}&\|\phi^{(n)} \cdot \Phi^{(n)} - \delta \cdot P_{\mathcal{B}^{(n)}}\Sigma_{XY}'P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}'\|_F \\ \leq&\|\phi^{(n)}\left(\Phi^{(n)} - P_{\mathcal{B}^{(n)}}\Sigma_{XY}'P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}'\right)\|_F \\ &+ \|\left(\phi^{(n)} - \delta\right) \cdot P_{\mathcal{B}^{(n)}}\Sigma_{XY}'P_{\mathcal{A}^{(n)}}'P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}'\|_F \\ \to&0\end{aligned}$$

as $n \to \infty$.

Thus, by Weyl's Theorem for Hermitian matrices, for each $i = 1, 2, \ldots, m$ we have almost surely that

$$\left|\lambda_i\left[\phi^{(n)} \cdot \Phi^{(n)}\right] - \lambda_i\left[\delta \cdot \left(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P'_{\mathcal{B}^{(n)}}\right)' P_{\mathcal{A}^{(n)}}\Sigma_{XY}P'_{\mathcal{B}^{(n)}}\right]\right| \to 0$$

as $n \to \infty$, from which Lemma 4 follows by noting that $P_{\mathcal{B}^{(n)}}$ is symmetric. Note that we use $\lambda_i(C)$ to denote the ith decreasingly ordered eigenvalue of any matrix C. □

We prove Theorem 2 in the next proof.

Proof. Let δ be as defined in Lemma 4. Note that for any non-negative, bounded real sequences $\{a^{(n)}\}_{n=1}^{\infty}$ and $\{b^{(n)}\}_{n=1}^{\infty}$, it holds that $a^{(n)} - b^{(n)} \to 0$ if and only if $\sqrt{a^{(n)}} - \sqrt{b^{(n)}} \to 0$, as $n \to \infty$. Thus, by Lemma 4, and noting that the rank of $P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}$ is at most d, we have almost surely that, as $n \to \infty$,

$$\sum_{i=1}^{m} \sigma_i(\mathcal{Y}^{(n)}\mathcal{X}^{(n)'}) - \sqrt{\delta} \cdot \sum_{i=1}^{d} \sigma_i(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}) \to 0. \tag{2.7}$$

The expression in Equation 2.7 can be simplified by Equation 2.6 and Equation 2.4 into

$$\begin{aligned}
&2d - 2\sum_{i=1}^{m} \sigma_i(\mathcal{Y}^{(n)}\mathcal{X}^{(n)'}) - \left[2d - 2\sqrt{\delta}\sum_{i=1}^{d} \sigma_i(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}})\right] \\
&= \epsilon^2 - \left[2d - 2\rho\sum_{i=1}^{d}\left(\frac{1}{\frac{1}{d}\sum_{j=1}^{d}\sigma_j\left(\Sigma_{XY}\right)}\sigma_i\left(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}\right)\right)\right] \\
&= \epsilon^2 - \left[(1-\rho)\cdot 2d + \rho\cdot\sum_{i=1}^{d} 2\left(1 - \frac{1}{\frac{1}{d}\sum_{j=1}^{d}\sigma_j\left(\Sigma_{XY}\right)}\sigma_i\left(P_{\mathcal{A}^{(n)}}\Sigma_{XY}P_{\mathcal{B}^{(n)}}\right)\right)\right] \\
&= \epsilon^2 - \left[(1-\rho)\cdot 2d + \rho\cdot\eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})\right],
\end{aligned}$$

which establishes Theorem 2. □

By noting the last paragraph in the proof of Lemma 3, and substituting the respective covariance matrices of Section 2.3 into Theorem 2, we immediately have Theorem 1.

In the next two propositions, we prove two important inequalities for the weighted Hausdorff distance $\eth$ and the ρ defined in Theorem 2.

Proposition 5. *For* $\eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)})$ *as defined in Equation 2.4, it holds that* $0 \leq \eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)}) \leq 2d$.

Proof. The upper bound is trivial. To prove the lower bound, first we re-express Equation 2.4 as

$$\eth^2(\mathcal{A}^{(n)}, \mathcal{B}^{(n)}) = \frac{2}{\frac{1}{d}\sum_{j=1}^{d} \sigma_j\left(\Sigma_{XY}\right)} \sum_{i=1}^{d} \left(\sigma_i(\Sigma_{XY}) - \sigma_i\Big(P_{\mathcal{A}^{(n)}} \Sigma_{XY} P_{\mathcal{B}^{(n)}}\Big)\right), \qquad (2.8)$$

and we show that each summand in the summation of Equation 2.8 is non-negative. Indeed, for any $S \in \mathbb{R}^{m \times m}$ and $i = 1, 2, \ldots, n$, we claim that $\sigma_i(S \cdot P_{\mathcal{A}^{(n)}}) \leq \sigma_i(S)$ and $\sigma_i(P_{\mathcal{A}^{(n)}} S) \leq \sigma_i(S)$, proved as follows:

Say $P_{\mathcal{A}^{(n)}} = QEQ^{'}$ is such that $Q \in \mathbb{R}^{m \times m}$ is orthogonal and E is diagonal with 1's and 0's on its diagonals. Then

$$\begin{aligned}
\sigma_i^2(S \cdot P_{\mathcal{A}^{(n)}}) &= \lambda_i(P_{\mathcal{A}^{(n)'}} S^{'} S P_{\mathcal{A}^{(n)}}) \\
&= \lambda_i(QEQ^{'} S^{'} S Q E Q^{'}) = \lambda_i(EQ^{'} S^{'} S Q E) \\
&\leq \lambda_i(Q^{'} S^{'} S Q) = \lambda_i(S^{'} S) = \sigma_i^2(S).
\end{aligned}$$

The inequality holds by the Interlacing Theorem for Hermitian matrices.

By a similar argument we have $\sigma_i(P_{\mathcal{A}^{(n)}} S) \leq \sigma_i(S)$; and applying these in succession yields that $\sigma_i(P_{\mathcal{A}^{(n)}} \Sigma_{XY} P_{\mathcal{B}^{(n)}}) \leq \sigma_i(\Sigma_{XY})$. □

Proposition 6. *For ρ defined in Theorem 2, it holds that $0 \leq \rho \leq 1$.*

Proof. Let $\Sigma_{XY} = U \Lambda V'$ be the singular value decomposition; i.e. $U, V \in \mathbb{R}^{m \times m}$ are orthogonal, and $\Lambda \in \mathbb{R}^{m \times m}$ is the diagonal matrix consisting of its non-increasing singular values. Define $M \in \mathbb{R}^{2m \times 2m}$ by

$$M = \begin{bmatrix} U' & 0_m \\ 0_m & V' \end{bmatrix} \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma'_{XY} & \Sigma_Y \end{bmatrix} \begin{bmatrix} U & 0_m \\ 0_m & V \end{bmatrix} = \begin{bmatrix} U'\Sigma_X U & \Lambda \\ \Lambda & V'\Sigma_Y V \end{bmatrix}$$

where $0_m \in \mathbb{R}^{m \times m}$ is the matrix of zeros.

A covariance matrix is always positive semi-definite, thus M is positive semi-definite as well as all of its principal sub-matrices. For each $j = 1, 2, \ldots, d$, the two-by-two sub-matrix consisting of the jth and $(j+m)$th rows and columns of M has non-negative diagonals and a non-negative determinant, thus $(U'\Sigma_X U)_{jj}(V'\Sigma_Y V)_{jj} \geq (\Lambda_{jj})^2$, i.e.

$$\sigma_j(\Sigma_{XY}) \leq \sqrt{\left(U'\Sigma_X U\right)_{jj}} \cdot \sqrt{\left(V'\Sigma_Y V\right)_{jj}}. \tag{2.9}$$

Now, summing Equation 2.9 over $j = 1, 2, \ldots, d$ and applying the Cauchy-Schwartz inequality to the right-hand side of Equation 2.9, we further obtain

$$\sum_{j=1}^{d} \sigma_j(\Sigma_{XY}) \leq \sqrt{\sum_{j=1}^{d} \left(U'\Sigma_X U\right)_{jj}} \cdot \sqrt{\sum_{j=1}^{d} \left(V'\Sigma_Y V\right)_{jj}}. \tag{2.10}$$

Because the vector of its diagonals always majorizes the vector of its eigenvalues for any Hermitian matrix, it follows that

$$\sum_{j=1}^{d} \left(U' \Sigma_X U \right)_{jj} \leq \sum_{j=1}^{d} \lambda_j (U' \Sigma_X U) = \sum_{j=1}^{d} \sigma_j (\Sigma_X), \tag{2.11}$$

similarly for Σ_Y and V.

Thus Proposition 6 directly follows from Equation 2.10 and Equation 2.11. $\square$

Chapter 3

Generalized Canonical Correlation Analysis for Classification

3.1 Introduction

In our previous chapter, we show that separate projection can cause the incommensurability phenomenon and be harmful for later inference; and joint projection like canonical correlation analysis (CCA) does not suffer the same phenomenon.

In this chapter we introduce a classification task, and prove that joint projection can help the classification error. Specifically, we show that the classification error can always be improved when more and more data sets are added to do generalized CCA, assuming the extra data sets satisfy certain sufficient condition. The background information is in Section 3.2. The necessary prerequisites are discussed in Section 3.3.

The sufficient condition and the following theorems are shown in Section 3.4. Some discussions are offered in Section 3.5 to relate the results to practical scenarios such as high-dimensional data and functional data, in addition to the classical multivariate setting the theorems are based on. Our theoretical results are illustrated via simulations, as well as a real data experiment on Wikipedia documents, in Section 3.6. All proofs are put into Section 3.7, including brief comments to elaborate on the sufficient conditions.

Note that this chapter is based on the paper [18].

3.2 Background

Let $(X, Y) \sim F_{XY}$ be an $\mathbb{R}^m \times \{1, \ldots, K\}$ random pair, where X is the feature vector and Y is the class label. In statistical pattern recognition (see, e.g., [36], [37]) one seeks a classifier $g : \mathbb{R}^m \to \{1, \ldots, K\}$ such that the probability of misclassification $L(g) = P\{g(X) \neq Y\}$ is acceptably small. Because modern data sets are often multi-dimensional, the feature vector X is assumed to be a multivariate random variable of dimension m and it is often beneficial to carry out the classification in some lower dimension d $(1 \leq d < m)$ as m is usually large. Therefore dimension reduction is applied to first embed X from $\mathbb{R}^m$ to $\mathbb{R}^d$, prior to subsequent classification.

Herein we consider only linear projections, which are commonly used and are the foundation for many nonlinear methods. We denote a linear projection from $\mathbb{R}^m$ to

$\mathbb{R}^d$ by an $m \times d$ matrix A; then $A'X$ (the $'$ sign denotes transpose) is the projected feature vector in $\mathbb{R}^d$. It follows that the classification error for a given classifier g (whose domain is $\mathbb{R}^d$ from now on) is $L_A = P\{g(A'X) \neq Y\}$.

Given a distribution F_{XY}, a classifier g, and a non-empty set of linear projections $\mathcal{A}$, we define an optimal projection $A^* \in \arg\min_{A\in\mathcal{A}}\{L_A\}$ and denote the corresponding minimum error as L_{A^*}. The set $\mathcal{A}$ and the existence of A^* are discussed in Section 3.3 and Assumption 1. Roughly speaking, L_{A^*} is the minimum error one can hope to achieve by choosing A cleverly among linear projections.

Assuming that the classifier g is specified, the crucial step is to choose the dimension reduction method. If we have only X available as the feature vector, then PCA (Principal Component Analysis) [1] is a natural choice, which is applied for classification in [38]. On the other hand, if there is an auxiliary feature Z_1 of dimension m_1 available, that is, $(X, Z_1, Y) \sim F_{XZ_1Y}$ on $\mathbb{R}^m \times \mathbb{R}^{m_1} \times \{1, \ldots, K\}$, then CCA (Canonical Correlation Analysis) [6] is applicable on the pair (X, Z_1) to derive the projection A, which is used in [33]. In general, if there are S auxiliary features $\{Z_s \in \mathbb{R}^{m_s}, s = 1, \ldots, S\}$ (we always assume $1 \leq d \leq \min\{m, m_1, \ldots, m_S\}$), then GCCA (Generalized Canonical Correlation Analysis) [39] is applicable on $(X, Z_1, \cdots, Z_S)$ to derive A based on X and the auxiliary features $\{Z_s\}$.

Note that our classification task remains the same, so that at the classification step we observe only X but not $\{Z_s\}$; and so by "GCCA/CCA is applicable" we mean "GCCA/CCA can be used to derive the projection matrix A for use in the

classifier $g(A'X)$". Furthermore, although CCA is a special case of GCCA, for clarity purposes we shall assume that GCCA uses at least two auxiliary features whenever GCCA is compared to CCA. If we consider those auxiliary features as extra data sets available for use, GCCA can make use of additional data sets compared to CCA, but we do not know whether these additional data sets will allow GCCA to outperform CCA. At this moment, we should also point out that another popular approach combines GCCA/CCA into the supervised learning step explicitly as a classification rule [40], [41], [42], which is empirically more suitable if classification is the only purpose; while in our setting we first apply GCCA/CCA to project the data, followed by the supervised learning step based on the projected data and known labels, which is a more general and more classical view in exploring given data and can be followed by other inference tasks such as testing, clustering, classification, etc. These two approaches are not in conflict with each other: one may first apply GCCA/CCA to project the data without the labels, followed by classification using supervised CCA (which in fact is equivalent to linear discriminant analysis in the two-class case [40]).

The above setting leads to the following questions. Does GCCA perform better than CCA in classification when using additional auxiliary features? From an application point of view, do additional data sets help in the later classification task, and what type of data sets should be included as auxiliary features in deriving the projection? It turns out the answer is not simple. We consider these questions theoretically, by deriving conditions on the auxiliary features that imply the superiority of GCCA.

Let us say the joint feature $(X, Z_1, \cdots, Z_S) \sim F_{S+1}$, and a projection matrix A derived from GCCA/CCA using X and s auxiliary features is denoted by A_{s+1}. Our main objective is to derive sufficient conditions on F_3 such that if $\max\{L_{A_2}\} = L_{A^*}$, then $L_{A_3} = L_{A^*}$, as well as sufficient conditions such that $L_{A^*} = L_{A_3} < \min\{L_{A_2}\}$; and their generalizations to F_{S+1} with arbitrary $s \geq 2$. (Note that when there are two auxiliary features, A_2 may come from applying CCA to either (X, Z_1) or (X, Z_2); hence the 'max' and 'min'.) Equivalently, the objective is to demonstrate that additional data sets can be useful for the classification task when conditions are satisfied.

3.3 Preliminaries

Given two auxiliary features Z_1 and Z_2, the joint distribution of (X, Z_1, Z_2) is denoted by $F_3 \in \Omega_3$, where Ω_3 is a family of multivariate distributions on $\mathbb{R}^{(m+m_1+m_2)}$. The overall covariance matrix of F_3 is denoted by

$$\Sigma_{F_3} = \begin{bmatrix} \Sigma_X & \Sigma_{XZ_1} & \Sigma_{XZ_2} \\ \Sigma'_{XZ_1} & \Sigma_{Z_1} & \Sigma_{Z_1Z_2} \\ \Sigma'_{XZ_2} & \Sigma'_{Z_1Z_2} & \Sigma_{Z_2} \end{bmatrix} \in \mathbb{R}^{(m+m_1+m_2)\times(m+m_1+m_2)}.$$

The overall covariance matrix, along with the individual Σ_X, Σ_{Z_1} and Σ_{Z_2}, are all assumed finite and positive semi-definite with rank no less than d.

We can consider GCCA/CCA either with the population covariances or with the sample covariances. For our theoretical analysis we consider the population covariances directly, while in the numerical section we use the sample covariances, which

are asymptotically equivalent in the classical multivariate setting under standard regularity conditions [43].

Identifying the CCA projection $A_2 = A_2(X, Z_1)$ can be approached as the problem of finding two sets of unit-length canonical vectors $\{a_i\}$ and $\{b_i\}$ to maximize the correlation between $a_i^{'}X$ and $b_i^{'}Z_1$ for each $i = 1, \ldots, d$. (The size of a_i is $m \times 1$ and the size of b_i is $m_1 \times 1$.) That is, we wish to identity

$$\arg\max_{a_i, b_i} \rho_{\{a_i^{'}X, b_i^{'}Z_1\}} = \frac{a_i^{'}\Sigma_{XZ_1}b_i}{\sqrt{a_i^{'}\Sigma_X a_i}\sqrt{b_i^{'}\Sigma_{Z_1}b_i}}, \tag{3.1}$$

subject to the *uncorrelated constraints*

$$\rho_{\{a_i^{'}X, a_j^{'}X\}} = \frac{a_i^{'}\Sigma_X a_j}{\sqrt{a_i^{'}\Sigma_X a_i}\sqrt{a_j^{'}\Sigma_X a_j}} = 0$$

$$\text{and} \quad \rho_{\{b_i^{'}Z_1, b_j^{'}Z_1\}} = \frac{b_i^{'}\Sigma_{Z_1} b_j}{\sqrt{b_i^{'}\Sigma_{Z_1} b_i}\sqrt{b_j^{'}\Sigma_{Z_1} b_j}} = 0, \forall j < i.$$

Then the $m \times d$ matrix $A_2 = [a_1, \ldots, a_d]$ is the CCA projection matrix for X, and $A_2^{'}X \in \mathbb{R}^d$ is the projected feature vector. Alternatively, a different $A_2 = A_2(X, Z_2)$ can be identified. Note that the arguments to A_2 – (X, Z_1) or (X, Z_2) – represent the choice of auxiliary features, and will be suppressed if the choice is clear or irrelevant in the context.

To identify the GCCA projection A_3 based on (X, Z_1, Z_2), we are looking for three sets of unit-length canonical vectors $\{a_i\}$, $\{b_i\}$ and $\{c_i\}$ as follows:

$$\begin{gathered} \arg\max_{a_i, b_i, c_i} (\rho^r_{\{a_i^{'}X, b_i^{'}Z_1\}} + \rho^r_{\{b_i^{'}Z_1, c_i^{'}Z_2\}} + \rho^r_{\{a_i^{'}X, c_i^{'}Z_2\}}) \\ \text{subject to } \rho_{\{a_i^{'}X, a_j^{'}X\}} = \rho_{\{b_i^{'}Z_1, b_j^{'}Z_1\}} = \rho_{\{c_i^{'}Z_2, c_j^{'}Z_2\}} = 0, \ \forall j < i, \end{gathered} \tag{3.2}$$

where the exponent r in the GCCA formulation (3.2) indicates the specific GCCA criterion. A common practice is to set $r = 1$ or 2, which maximizes either the sum of correlations or the sum of squared correlations [39]. Then $A_3 = [a_1, \ldots, a_d]$ is the desired GCCA projection. In general, given F_{S+1} we can derive the GCCA projection A_{s+1} for any $1 \leq s \leq S$, and CCA is merely a special case for $s = 1$. Because our results are shown to hold for any $r \geq 1$, we implicitly take $r = 1$ unless mentioned otherwise.

Given Σ_X, we shall call an $m \times d$ matrix $A = [a_1, \ldots, a_d]$ a "potential" GCCA projection if and only if its columns $\{a_i\}$ are of unit-length and satisfy the uncorrelated constraints. The set containing all potential GCCA projections is denoted by $\mathcal{A} = \{A|\ \rho_{\{a_i'X, a_j'X\}} = 0\ \forall i \neq j \text{ and } \|a_i\| = 1\ \forall i\}$. As a different choice of auxiliary features yields a different projection, we denote the set containing the GCCA projections A_3 by $\mathcal{A}_3$ and the set containing all CCA projections A_2 by $\mathcal{A}_2$, as well as the set $\mathcal{A}_{s+1}$ in general. Clearly the elements of $\mathcal{A}_{s+1}$ as well as $\mathcal{A}$ depend on Σ_X. Note that the PCA projection is also an element of $\mathcal{A}$, but this is not of our concern in this chapter. An important special case: $\mathcal{A}$ represents the Stiefel manifold [44] (containing all orthogonal projections onto dimension d linear subspaces) when Σ_X is a multiple of the identity.

Note that the original GCCA/CCA algorithm does not require the norm of a_i to be the same for all i. We choose them to be unit-length consistently in order to avoid scaling issues in the classification step (alternatively, it is a common practice to set

$a_i^{'}\Sigma_X a_i = 1$ for all i, which is equivalent for our purposes). Also note that the choice of the GCCA/CCA projections can be arbitrary. For example, let Σ_X and Σ_{Z_1} be identity matrices and all the singular values of Σ_{XZ_1} be the same; then $A_2(X, Z_1)$ can be chosen arbitrarily in the Stiefel manifold $\mathcal{V}_{d,m}$. In this case A_2 has $md - \frac{d^2+d}{2}$ degrees of freedom, where md comes from the dimension freedom by repeating singular values and $\frac{d^2+d}{2}$ comes from the unit-length requirement and uncorrelated constraints. But if Σ_{XZ_1} does not have repeating singular values, A_2 represents a fixed subspace and has $\frac{d^2-d}{2}$ degrees of freedom, which is implied by the fact that two $m \times d$ matrices A and B represent the same subspace if and only if $AA^{'} = BB^{'}$. The same phenomenon applies for any GCCA projection A_{s+1}.

Returning to the classification problem: given a classifier $g : \mathbb{R}^d \to \{1, \ldots, K\}$ for the low-dimensional feature vector $A^{'}X$, the error L_A may differ for different $A \in \mathcal{A}$. Clearly $\mathcal{A}$ is compact for finite Σ_X and $\{L_A | A \in \mathcal{A}\}$ is bounded between $[0, 1]$, but an optimal low-dimensional projection (with respect to the classification error) is not guaranteed to exist. We make the following assumption to avoid non-existence:

Assumption 1. *Given a classifier g, we assume for the theory in the sequel that an optimal projection $A^* = \arg\min_{A \in \mathcal{A}}\{L_A\}$ exists for any finite Σ_X of rank at least d.*

For example, if the class-conditional distributions $F_{X|Y=k}$ admit probability density functions $f_{X|Y=k}$ for $k = 1, \ldots, K$, then the assumption always holds. (In this case L_A is continuous with respect to A, and thus $\{L_A | A \in \mathcal{A}\}$ is compact and admits a minimum.)

By this assumption, the minimum error L_{A^*} always exists and it follows that $L_{A_{s+1}} \geq L_{A^*}$ always holds for any s. Note that the optimal projection A^* need not be unique, since the existence suffices for our purposes. Now we are able to define the notion that GCCA improves CCA using L_{A^*}.

Definition 1. Assuming the existence of A^*, we say GCCA improves CCA within a family of distributions Ω_3 if and only if $\{F_3 \in \Omega_3 | L_{A_2} = L_{A^*},\ \forall A_2 \in \mathcal{A}_2\} \subset \{F_3 \in \Omega_3 | L_{A_3} = L_{A^*},\ \forall A_3 \in \mathcal{A}_3\}$.

In general, we say the set of GCCA projections $\mathcal{A}_{s+1}$ improves the set of GCCA projections $\mathcal{A}_{t+1}$ within Ω_{S+1} $(1 \leq s, t \leq S)$ if and only if $\{F_{S+1} \in \Omega_{S+1} | L_{A_{t+1}} = L_{A^*},\ \forall A_{t+1} \in \mathcal{A}_{t+1}\} \subset \{F_{S+1} \in \Omega_{S+1} | L_{A_{s+1}} = L_{A^*},\ \forall A_{s+1} \in \mathcal{A}_{s+1}\}$. (Here the notation "$\subset$" indicates proper subset.)

Put in words, suppose GCCA improves CCA within Ω_3. Then the optimality of the CCA projections implies the optimality of the GCCA projection, and there exists F_3 such that the GCCA projection is optimal while at least one of the CCA projections is not. Such improvement implies that additional data sets should be used, though it is not equivalent to $L_{A_3} \leq L_{A_2}$.

If Ω_3 includes every possible multivariate distribution, then GCCA fails to improve CCA. For example, if Z_1 and Z_2 are both positively correlated to X but Z_1 and Z_2 are negatively correlated, then it might happen that A_2 is optimal while A_3 is not. Hence it is not always a good idea to incorporate additional auxiliary features, and we shall look for a family Ω_3 imposing certain relationships among X and $\{Z_s\}$ such

that GCCA is guaranteed to improve CCA.

First, we transform X by centering and whitening, so that the population mean is zero and the population covariance matrix becomes the identity matrix. Then $\mathcal{A}$ consists of orthogonal projections onto dimension d linear subspaces, and there exists an orthogonal matrix such that the feature vector can be rotated to guarantee A^* is equivalent to the subspace $\mathbb{R}^d$ spanned by the first d coordinate axes. We denote the transformed random variable by $\tilde{X} = H_X(X - E(X))$, where $E(X)$ is the expectation for centering and H_X is a non-singular $m \times m$ matrix for whitening and rotation. Since the optimal projection for $\tilde{X}$ is spanned by the first d coordinate axes, the form of $\tilde{X}$ based on the class label $Y = \{1, \ldots, K\}$ can be expressed as:

$$\tilde{X} = H_X(X - E(X)) \overset{law}{=} \begin{bmatrix} U_1\mathbf{1}_1 + U_2\mathbf{1}_2 + \cdots + U_K\mathbf{1}_K \\ \\ W \end{bmatrix}, \tag{3.3}$$

where $\mathbf{1}_k$ is the class label indicator taking value k with probability p_k and $\sum_{k=1}^{K} p_k = 1$, each $U_k \in \mathbb{R}^d$ is the marginal distribution of $\tilde{X}$ under class k, and $W \in \mathbb{R}^{m-d}$ is the "irrelevant" marginal of $\tilde{X}$. By the above transformation it holds that $E(W) = 0_{(m-d)\times 1}$ and $E(WW^{'}) = I_{(m-d)\times(m-d)}$, where I denotes the identity matrix. Clearly H_X always exists, and there are multiple choices for H_X if A^* is not unique. Now we impose our conditions on F_{S+1} and define what we call the similar family.

3.4 Main Results

Definition 2. We say the family of distributions Ω^*_{S+1} is *the similar family* if and only if it includes every F_{S+1} such that $(X, Z_1, \cdots, Z_S) \sim F_{S+1}$ satisfies the following conditions:

Condition (1): For each A^*, there exists non-singular matrices $H_X \in \mathbb{R}^{m\times m}$ and $H_{Z_s} \in \mathbb{R}^{m_s \times m_s}$ for all $s = 1, \ldots, S$, such that Equation (3.3) holds and there exist non-negative scalars q_{sk} with

$$\tilde{Z}_s = H_{Z_s}(Z_s - E(Z_s)) \overset{law}{=} \begin{bmatrix} q_{s1}U_1\mathbf{1}_1 + q_{s2}U_2\mathbf{1}_2 + \cdots + q_{sK}U_K\mathbf{1}_K + e_s \\ \\ W_s \end{bmatrix}, \quad (3.4)$$

where e_s represents independent noise and $W_s \in \mathbb{R}^{m_s - d}$. Note that unlike H_X, H_{Z_s} need only be non-singular and Z_s are not necessarily whitened and rotated.

Condition (2): $E(U_k U_k^{'}) = I$, and U_k is uncorrelated with W and W_s, for all $k = 1, \ldots, K$ and $s = 1, \ldots, S$.

Condition (3): $\sigma_1(E(W_s W_t^{'})) \leq \sigma_1(E(WW_s^{'}))\sigma_1(E(WW_t^{'}))$ for all $1 \leq s \neq t \leq S$, where we denote $\sigma_i(\Sigma)$ as the ith largest singular value for any matrix Σ henceforth.

Condition (4): $(q_{sk_1} - q_{sk_2})(q_{tk_1} - q_{tk_2}) > 0$ for all $1 \leq s < t \leq S$ and $k_1, k_2 = 1, \ldots, K$; namely the ordering of coefficients q_{sk} is consistent throughout Z_s.

The purpose of condition (1) is to guarantee that the marginal distribution restricted to A^* of every transformed auxiliary feature under each class is a scalar

multiple of the corresponding marginal of $\tilde{X}$ plus error. The possible non-uniqueness of A^* is (mostly) avoided by requiring (1) to hold for any A^*, though the transformation matrices and respective scalars probably differ under different A^*. Condition (2) is to simplify the analysis, without which the proof is much more complex. Given conditions (1) and (2), conditions (3) and (4) are technical conditions used in the proof, implying certain relationships among features. Interpreted by words, condition (3) implies the "noisy" dimensions (where W and W_s live in) among the auxiliary features should be less related, while condition (4) implies the "signal" dimensions (where U_k lives) among the auxiliary features should be more related. In this case GCCA is more likely to extract information from the "signal" dimensions, for which utilizing additional data sets is likely to improve the classification error. As we will see in the numerical experiments, this interpretation is useful for judging qualitatively whether additional data sets should be included, even if A^* is unknown or condition (2) is not satisfied. And we will provide additional comments at the end of the proof section to discuss the magnitude of q_{sk} and its potential impact on the sufficient conditions and model selection.

Theorem 7. *GCCA improves CCA in the similar family* Ω_3^*.

Therefore it is beneficial to use the GCCA projection A_3 within the similar family Ω_3^*, whose conditions are sufficient but not necessary for GCCA to improve CCA. Equivalently, deriving the projection using additional data sets helps the classification task when the sufficient conditions are satisfied.

Furthermore, the similar family can be decomposed into three disjoint subsets as follows: $\Omega_3^* = \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} = L_{A_3} = L_{A^*}\} \cup \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\} \cup \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A^*}$ and $L_{A_3} > L_{A^*}\}$, with all the subsets shown to be non-empty and proper in the proof (we can also replace all the 'max' by 'min'). Specifically, if the optimal A^* is known (which may be difficult in practice), then one can check which subset a given $F_3 \in \Omega_3^*$ belongs to according to Inequality (3.5) and Inequality (3.6) in the proof below. When the distribution lies in the first or the second subset above, the GCCA projection performs no worse than the CCA projections, and adding a "qualified" additional dataset yields better classification result.

It is natural to consider a generalization to Ω_{S+1}^* because there may be many additional data sets satisfying the conditions. Indeed we have an easy generalization of the above theorem.

Corollary 1. *For any $S \geq S' \geq 2$, the set of GCCA projections $\mathcal{A}_{S'+1}$ improves the set of CCA projections $\mathcal{A}_2$ in the similar family Ω_{S+1}^*.*

Under a simplified setting, we can also show that the set of GCCA projections continue to improve when additional auxiliary features are included in deriving the projections. This means in the context of the similar family, additional data sets will always improve the performance in the classification task.

Corollary 2. *Let us replace condition (4) by a simplifying condition (4'): $W_s = W_t$ and $q_{sk} = q_{tk}$ for all $1 \leq s, t \leq S$. Namely the auxiliary features follow the same*

distribution for $s = 1, \ldots, S$.

Then for any $S \geq S^{'} \geq 2$, *the set of GCCA projections* $\mathcal{A}_{S^{'}+1}$ *always improves the set of GCCA projections* $\mathcal{A}_{S^{'}}$ *in the similar family* Ω^{*}_{S+1}.

3.5 Discussions

Since our analysis is carried out on the population covariance instead of the sample covariance, our results so far rely on the fact that the sample covariance converges to the population covariance as dimension reduction methods including GCCA/CCA are mostly carried out on the sample data. Let us provide some justifications for the high-dimensional data case, where the dimension m is large when compared to the number of training observations n' such that the covariance convergence is not guaranteed.

For high-dimensional data, if the sample covariance is still close to the true covariance with high probability as discussed in [45] and [46], then our results still apply and GCCA improves CCA in the similar family with high probability. Otherwise our conditions in Definition 2 cannot be directly used to justify the GCCA/CCA behavior on sample covariances of high-dimensional data. However, one may heuristically claim that if GCCA is better than CCA in the population model for the classification task, then GCCA is expected to be better than CCA for the sample data: Since the classification error is actually a function of the data, if $L_{A_3} < L_{A_2}$ for A_2 and

A_3 derived from the population model, then at a suitable level of n'/m we can have $Prob\{L_{A_3} < L_{A_2}\} > 0.5$ for A_2 and A_3 derived from the sample data, because this probability converges to 1 in the classical multivariate setting where $n'/m \to \infty$. (A point of interest is to derive the minimum level n'/m, which may depend on the classifier we use. For our simulations on the synthetic data generated within the similar family, it seems the minimum level is no larger than 1 in order for GCCA to be better than CCA.)

In practice one rarely applies CCA directly on data of very high dimension with $m > n$. Often one opts to use kernel CCA [47], [48], sparse CCA [49], [50] or functional CCA [51], [52] to deal with noisy high-dimensional data, assuming that the data intrinsically lives in some low-dimensional linear subspace. For example, instead of working on $(X, Y) \in \mathbb{R}^m$ where m is very large, kernel/functional CCA works on $(f(X), g(Y))$ by assuming appropriate f and g exist for nonlinear/functional data. But the analysis of sparse/functional CCA will be quite different and difficult when penalty terms are introduced in the constraints, which requires numerical methods to solve and gives different GCCA/CCA transformations that cannot be efficiently expressed in matrix notation.

Another aspect worth noting is that a similar conclusion may be reached for clustering. This is because GCCA makes it easier to find the optimal subspace than CCA under the same conditions, as long as one is able to define an optimal subspace A^* in terms of some clustering algorithm with respect to a specific performance index.

However, we do not pursue this direction here because it is more challenging to evaluate clustering performance than classification performance.

Furthermore, since GCCA/CCA does not make use of label information in the dimension reduction step, it is natural to compare with some existing algorithms such as p-LDA (penalized linear discriminant analysis) [40], [53] and ℓ1-SVM (1-norm support vector machine) [54], [55], which make use of labels and may work for data of high/unknown dimensions. Even though we will include their classification results in the numerical section for benchmark purposes, our target is not to find the best method for a given dataset. In addition to being more appropriate for an exploratory task, there are other reasons that applying unsupervised dimension reduction methods first is more favorable than doing supervised dimension reduction directly, e.g., it is easier and faster to use unsupervised dimension reduction for real data, it may be slow and difficult to choose a suitable penalty term in p-LDA, the data before dimension reduction may not have access to the labels or may be different from the data on which we perform classification as in the transfer learning task [56], etc.

At last, the choice of projection dimension d is crucial for the classification (or any inference) performance, especially when working with real data of unknown true dimension. There are a number of papers on dimension choice for projecting a single dataset [57], [58] but not for multiple correlated data sets, which may be an interesting point to pursue. Still, our results are always valid no matter the choice of d, which

means GCCA improves CCA for any d when conditions are satisfied.

3.6 Numerical Experiments

To investigate the performance of the GCCA/CCA projections in classification, we present both numerical simulations and a real data experiment. We use sample covariances to derive the GCCA projections with the GCCA algorithm implemented according to [42] (though no covariance matrix regularization is required in our experiments in contrast to their RGCCA algorithm; and we apply Gram-Schmidt to all output vectors in the iteration of the algorithm to enforce the uncorrelated constraints of all the canonical vectors), and the usual LDA as our main classification rule for the following supervised learning. Whenever applicable, we also include p-LDA and ℓ1-SVM classification results based on the single dataset to compare with the LDA classification results based on the GCCA/CCA projected dataset. Note that our previous numerical work illustrating GCCA improvement under kNN (k-nearest neighbor) classifier is available in [35].

3.6.1 Numerical Simulations

We start with four random variables $U_1, U_2 \in \mathbb{R}^3$ and $V_1, V_2 \in \mathbb{R}^6$ all independently normally distributed. The parameters are set as follows: $E(U_1U_1') = E(U_2U_2') = I_{3\times 3}$, $E(U_1) = -E(U_2) = 0.2_{3\times 1}$, $E(V_1V_1') = E(V_2V_2') = 0.5I_{6\times 6}$, $E(V_1) = E(V_2) = 0_{6\times 1}$.

The three random variables $X, Z_1, Z_2 \in \mathbb{R}^9$ are constructed as follows:

$$X \overset{law}{=} \begin{bmatrix} U_1\mathbf{1}_1 + U_2\mathbf{1}_2 \\ V_1 + V_2 \end{bmatrix},$$
$$Z_1 \overset{law}{=} \begin{bmatrix} 0.6U_1\mathbf{1}_1 + 0.4U_2\mathbf{1}_2 + e_1 \\ V_1 + e_3 \end{bmatrix},$$
$$Z_2 \overset{law}{=} \begin{bmatrix} 0.6U_1\mathbf{1}_1 + 0.4U_2\mathbf{1}_2 + e_2 \\ V_2 + e_4 \end{bmatrix},$$

where $e_1, e_2 \overset{\text{i.i.d.}}{\sim} N(0, 0.75I_{3\times3})$, $e_3, e_4 \overset{\text{i.i.d.}}{\sim} N(0, 0.5I_{6\times6})$, $\mathbf{1}_1$ and $\mathbf{1}_2$ are class label indicators having equal probability. Using LDA, it is clear that at $d = 3$ the ideal optimal projection A^* uniquely represents the subspace spanned by the first d coordinate axes. Hence we can fit the joint distribution into Definition 2 with $d = 3$, such that $q_{11} = q_{21} = 0.6$, $q_{12} = q_{22} = 0.4$, $W = V_1 + V_2$, $W_1 = V_1 + e_3$, $W_2 = V_2 + e_4$, etc. This joint distribution satisfies the required conditions, so it belongs to Ω_3^*. Further, by checking Inequality (3.5) and Inequality (3.6) in the proof, the joint distribution is actually an element of the subset $\{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\} \in \Omega_3^*$. So we expect GCCA to outperform CCA when projected onto $\mathbb{R}^3$. Note that in this case we can explicitly calculate L^* for the population model, which is 36.45%.

For each Monte Carlo replicate, $n = 1500$ observations are generated for each random variable. That is, $\{x^{(1)}, \ldots, x^{(1500)}\}$ for X, $\{z_1^{(1)}, \ldots, z_1^{(1500)}\}$ for Z_1 and $\{z_2^{(1)}, \ldots, z_2^{(1500)}\}$ for Z_2. All data points are used to learn the GCCA/CCA projections respectively for $d = 3$. (One may instead derive the projections based on the training data only, which is asymptotically equivalent to deriving the projections from all the available data if the testing data is distributed the same as the train-

ing.) Then the first 1000 points generated from X are projected and used to train the classifier; the remaining 500 points are projected and used for classification error testing. The classification error is recorded separately for the CCA projections $A_2(X, Z_1)$ and $A_2(X, Z_2)$ and for the GCCA projections A_3, using both sum of correlation ($r = 1$) and sum of squared correlation ($r = 2$) criteria. The above is done for 500 Monte Carlo replications, and we show in Table 3.1 the average classification error and the average difference between the derived GCCA/CCA subspace and the optimal subspace for each projection (we use the Hausdorff distance [28] for the difference between subspaces). The average GCCA classification error is lower than that of CCA as expected, and is fairly close to the optimal error L^*. In this case the average errors using the p-LDA and $\ell 1$-SVM are 37.37% and 36.50% respectively (the penalty terms are always chosen based on cross-validation for the best performances and benchmark purposes). Note that the standard deviations for the average errors of all the methods are within 0.3%, and those for the distance of the subspaces are within 0.002, which are the same for all the later simulations. Also note that the distances of the subspaces are not expected to be 0, because the A^* we use is the ideal optimal subspace for the population model and different from the optimal subspace for the sample data; but even so, it seems that the classification error is positively correlated to the distance of the subspaces.

To investigate the effect of higher dimension and less sample data, we repeat the same procedure three times, for $m = 20$ with $n = 1500$, $m = 50$ with $n = 1500$,

and $m = 50$ with $n = 75$ (50 points used for training and the remaining 25 used for testing). The settings are the same with $d = 3$ fixed, e.g., the dimensions of U_i stay at 3 but the dimensions of V_i are increased as m increases. The results are shown in Table 3.2, Table 3.3 and Table 3.4. A higher dimension or a smaller training size means the sample covariance does a worse job in approximating the population covariance, possibly making the differences between the derived GCCA/CCA subspace and the optimal subspace larger as m increases and/or n decrease; but still GCCA is better than CCA for the classification task in all the tables, reflecting our heuristic argument in the discussion section. This time the average errors using the p-LDA and ℓ1-SVM are 39.01% and 39.27% at $m = 20$ with $n = 1500$, 38.91% and 38.81% at $m = 50$ with $n = 1500$, and 47.43% and 45.76% at $m = 50$ with $n = 75$, most of which turn out to be slightly better than using LDA on GCCA projected data throughout these simulations.

We also present another simulation to show that GCCA does not necessarily improve CCA at $m = 9$ with $n = 1500$, by replacing the auxiliary feature Z_2 by $Z_{2'} \overset{law}{=} \begin{bmatrix} 0.6U_1\mathbf{1}_1+0.4U_2\mathbf{1}_2+e_2 \\ V_1+e_4 \end{bmatrix}$. We re-generate all observations and carry out the same simulation steps. Although the auxiliary feature $Z_{2'}$ looks reasonably "similar" to X (differing from Z_1 only by noise), the joint distribution of $(X, Z_1, Z_{2'})$ does not satisfy condition (3) and GCCA does not improve CCA by checking the covariance structure explicitly. Interpreted by words, Z_1 and $Z_{2'}$ are too correlated in the "noisy" dimensions, hindering GCCA from recognizing the correct "signal" dimensions. The

projections	CCA on (X, Z_1)	CCA on (X, Z_2)	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	42.03%	41.89%	37.00%	38.16%
$\|A - A^*\|$	1.688	1.591	0.714	0.989

Table 3.1: GCCA Improves CCA in simulation at $m = 9, n = 1500$

projections	CCA on (X, Z_1)	CCA on (X, Z_2)	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	47.02%	46.18%	42.84%	44.19%
$\|A - A^*\|$	2.161	2.037	1.364	1.825

Table 3.2: GCCA Improves CCA in simulation at $m = 20, n = 1500$

projections	CCA on (X, Z_1)	CCA on (X, Z_2)	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	47.58%	46.02%	42.41%	44.31%
$\|A - A^*\|$	2.197	2.161	1.643	1.895

Table 3.3: GCCA Improves CCA in simulation at $m = 50, n = 1500$

projections	CCA on (X, Z_1)	CCA on (X, Z_2)	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	51.98%	51.60%	45.76%	49.98%
$\|A - A^*\|$	2.256	2.236	2.179	2.203

Table 3.4: GCCA Improves CCA in simulation at $m = 50, n = 75$

average simulated classification errors are shown in Table 3.5. In this case GCCA performs worse than CCA, which demonstrates that simply adding more data sets does not automatically yield a better result.

projections	CCA on (X, Z_1)	CCA on $(X, Z_{2'})$	GCCA $(r = 1)$	GCCA $(r = 2)$
average error (L_A)	41.34%	41.33%	46.86%	46.90%
$\|A - A^*\|$	1.545	1.537	2.009	2.018

Table 3.5: GCCA Fails to Improve CCA in simulation

3.6.2 Wikipedia Documents

The real data experiment applies GCCA/CCA to text document classification. The dataset is obtained from Wikipedia, an open-source multilingual web-based encyclopedia with millions of articles in more than 280 languages. In Wikipedia each article can be related to others in the same language, or articles in other languages with the same subject. Articles of the same subject in different languages are not necessarily exact translations of one another; it is very likely they are written by different people and their contents might differ significantly.

English articles within a 2-neighborhood of the English article "Algebraic Geometry" are collected, and the corresponding French articles of those English documents are also collected, which totals $n = 1382$ pairs of articles in English and French. Let $a_1^e, \ldots, a_{1382}^e$ denote the English articles and $a_1^f, \ldots, a_{1382}^f$ denote the French articles. All articles are manually labeled into 5 disjoint classes $(1 - 5)$ based on their topics, as shown in Table 3.6.

For the purposes of GCCA/CCA, first we need to embed each article onto the Euclidean space $\mathbb{R}^m$ by Multi-dimensional Scaling (MDS) [3], [5], [4]. MDS strives

topic	category	people	locations	date	math
class label	1	2	3	4	5
article number	119	372	270	191	430

Table 3.6: Wikipedia Dataset Topics

to give a Euclidean representation while approximately preserving the dissimilarities of the original data: given an $n \times n$ dissimilarity matrix $\Delta = [\delta_{ij}]$ for n observations with δ_{ij} being the dissimilarity measure between the ith and jth observation, MDS generates embeddings $x_i \in \mathbb{R}^m$ for the ith data point to preserve the dissimilarity among the objects pairs, i.e. $||x_i - x_j|| \approx \delta_{ij}$.

For our work two different types of dissimilarity measures are considered for English and French articles, giving four dissimilarity matrices of dimension 1382×1382: the graph topology dissimilarity matrix $\bar{\Delta}^e, \bar{\Delta}^f$ and the text content dissimilarity matrix $\hat{\Delta}^e, \hat{\Delta}^f$.

For the graph dissimilarities, $\bar{\Delta}^e$ and $\bar{\Delta}^f$ are constructed based on an undirected graph $G(V, E)$, where V represents the set of vertices of the 1382 Wikipedia documents, and E is the set of edges connecting those articles. There is an edge between two vertices if they are linked in Wikipedia. Then the entry $\bar{\Delta}^e(i, j)$ is calculated from the number of steps on the shortest path from document i to document j in G. For the English articles, $\bar{\Delta}^e(i, j) \in \{0, \ldots, 4, 6\}$, where 4 is the upper bound of the step number with any higher number setting to 6. For the French articles $\bar{\Delta}^f(i, j)$

	Graph Topology Dissimilarity	Text Content Dissimilarity
English articles $\{a_i^e\}$	$\{\bar{x}_i^e\}(GE)$	$\{\hat{x}_i^e\}(TE)$
French articles $\{a_i^f\}$	$\{\bar{x}_i^f\}(GF)$	$\{\hat{x}_i^f\}(TF)$

Table 3.7: Euclidean Embeddings ($\mathbb{R}^m$) for Wikipedia Articles

depends on the French graph connections, so it is possible that $\bar{\Delta}^f(i,j) \neq \bar{\Delta}^e(i,j)$. At the extreme end, $\bar{\Delta}^f(i,j) = \infty$ when a_i^f and a_j^f are not connected, and we set $\bar{\Delta}^f(i,j) = 6$ for $\bar{\Delta}^f(i,j) > 4$.

For the text dissimilarities, $\hat{\Delta}^e$ and $\hat{\Delta}^f$ are based on the text processing features for documents $\{a_i^e\}$ and $\{a_i^f\}$. Suppose $\mathbf{z}_i, \mathbf{z}_j$ are the feature vectors for the ith and jth English articles. Then $\hat{\Delta}^e(i,j)$ is calculated by the cosine dissimilarity $\hat{\Delta}^e(i,j) = 1 - \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$. For the experiment we consider the latent semantic indexing (LSI) features [59].

Once different dissimilarity matrices are constructed, the Euclidean space embeddings with $m = 50$ are obtained via MDS. The articles' embeddings are shown in Table 3.7. At first, English graph dissimilarity (GE) is the classification target, and others (GF, TE, TF) are treated as auxiliary features: all data points are used to learn the GCCA/CCA projections from $\mathbb{R}^m$ to $\mathbb{R}^d$ based on GE and a certain choice of auxiliary features, and the data points of GE are projected by the learned projections. Then 600 observations are randomly picked to train the classifier, with the remaining 782 documents used for classification error testing. We repeat 500 times to

calculate the average classification error, for every possible GCCA/CCA projection and various choice of d. The same procedure is repeated with the French graph dissimilarity (GF) being the classification target and the remaining being the auxiliary features. The full results for every possible projection are shown in Figure 3.1 for the classification of GE. For illustration purposes, two simplified plots are shown in Figure 3.2 for the classification of GE/GF, for which we omit most projections in order to better quantify the effects of increasing s (the number of chosen auxiliary features), i.e., only the best A_2 and A_3 are shown. Note that for comparison purposes the PCA projections are also included, and all the classification errors have standard deviations within 0.2%.

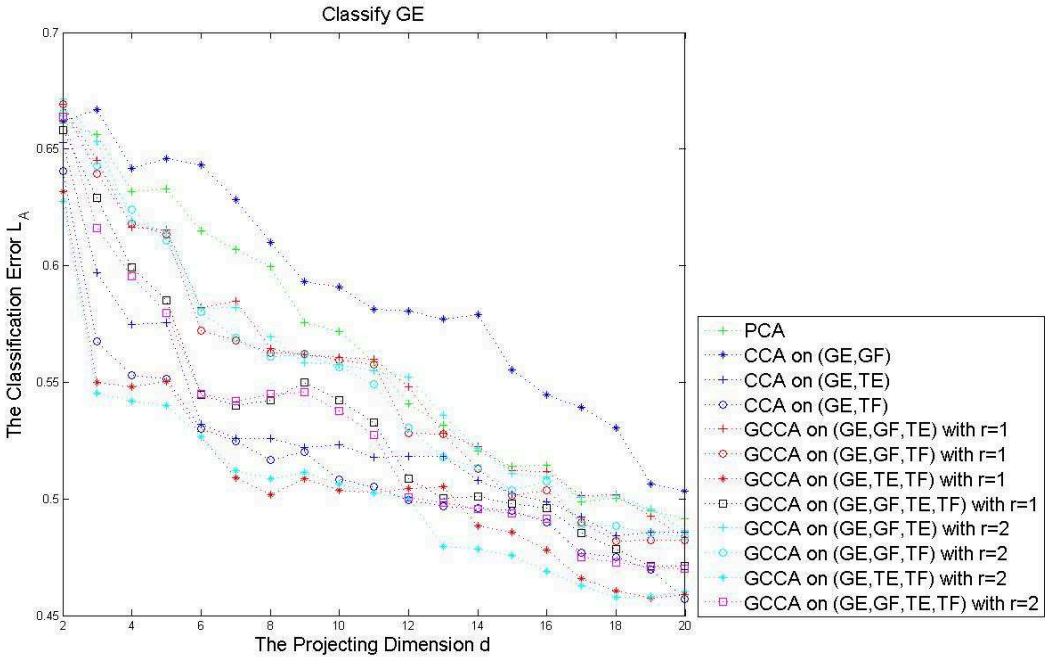
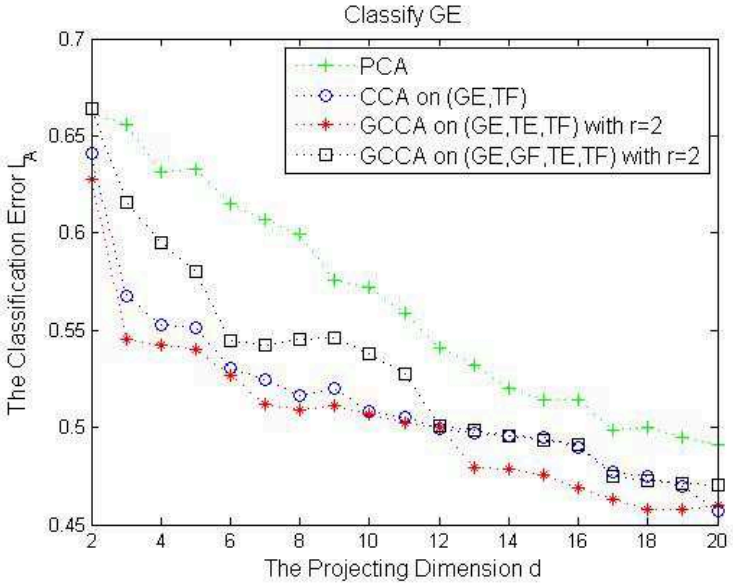
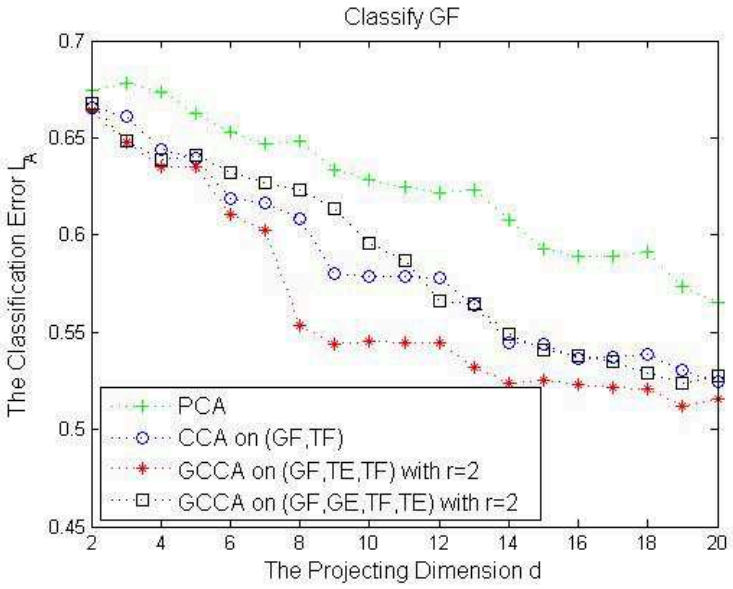


Figure 3.1: Classification Error for GE



(a)



(b)

Figure 3.2: Classification Error for GE/GF (simplified)

Based on Figure 3.2, we observe that for most choices of d the best GCCA projection A_3 admits a lower error than the best CCA projection A_2, and both of them are better than the PCA projection. The figure also illustrates the last paragraph of our discussion section, i.e., GCCA is expected to be better than CCA no matter the choice of projection dimension. However, it turns out that the GCCA projection A_4 does not yield the lowest error for classifying the Wikipedia data. This is not a surprise and tells that not all data sets should be included in this example, as one can judge from Figure 3.1 and our previous simulations that the choice of auxiliary features is crucial for the classification errors. For benchmark purposes, the average classification errors using p-LDA on the MDS-embedded data are 48.40% for GE and 56.65% for GF, which are slightly better than the average LDA errors using PCA projected data but worse than the average LDA errors using multiple data sets and the best GCCA/CCA projections at $d = 20$ in this experiment.

Unfortunately, one cannot easily check the joint distribution by Definition 2 like in the simulation part, because the optimal projection A^* is unknown for the Wikipedia data sets. Therefore in a real-world application, one must be cautious in adding a new dataset and/or choosing the best dimension. Both of these are difficult model selection problems in practice, which can be addressed by cross-validation as in this experiment. Still, the interpretation after Definition 2 is useful from a qualitative perspective. On one hand, the graph dissimilarities GE and GF are of questionable value because they depend on the Internet links, which may be erroneous. On the

other hand, the text dissimilarities TE and TF are much more faithful because they are extracted from the document contents, thus more likely to have commonality in certain "signal" dimensions. Therefore it is reasonable to believe that choosing a text dissimilarity is better than choosing a graph dissimilarity, which explains why the best A_2 and A_3 do not choose any graph dissimilarity as the auxiliary variable and why A_4 performs worse.

3.7 Proofs

3.7.1 Proof of Theorem 7 when $K = 2$ and $r = 1$

Proof. We consider $K = 2$ and $r = 1$ here (and generalize in the next proof), so the number of classes is two and the GCCA criterion is the sum of correlations.

If a projection A represents the same subspace as the optimal projection A^* (i.e., $AA' = A^*A^{*'}$), then A is optimal for classification such that $L_A = L_{A^*}$. For most parts it suffices to assume that A^* is unique (in the sense of representing the same subspace), which is justified towards the end of the proof.

In addition to the uniqueness of A^*, we also assume that $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices for $s = 1, 2$. This is also justified later, as we will show the theorem is invariant under proper transformations. Further, the expectations $E(X)$ and $E(Z_s)$ are treated as zeros throughout all proofs because the GCCA/CCA projections and the classification task are not affected.

Under the above assumptions, we have the following: the optimal projection A^* is spanned by the first d coordinate axes; any potential projection $A \in \mathcal{A}$ must be orthonormal and equivalent to an orthogonal projection onto a dimension d linear subspace; and the GCCA/CCA projections A_{s+1} are optimal if and only if $A_{s+1}A_{s+1}^{'} = A^*A^{*'}$.

Because all the pre-multiplication matrices are assumed to be identity matrices, together with conditions (1) and (2) in Definition 2 we have the covariance matrices

$$\Sigma_{XZ_1} = \begin{bmatrix} pq_{11}E(U_1U_1^{'}) + (1-p)q_{12}E(U_2U_2^{'}) & pE(U_1W_1^{'}) + (1-p)E(U_2W_1^{'}) \\ \\ pq_{11}E(WU_1^{'}) + (1-p)q_{12}E(WU_2^{'}) & E(WW_1^{'}) \end{bmatrix}$$

$$= \begin{bmatrix} (pq_{11} + (1-p)q_{12})I_{d\times d} & 0 \\ \\ 0 & E(WW_1^{'}) \end{bmatrix},$$

$$\Sigma_{XZ_2} = \begin{bmatrix} pq_{21}E(U_1U_1^{'}) + (1-p)q_{22}E(U_2U_2^{'}) & pE(U_1W_2^{'}) + (1-p)E(U_2W_2^{'}) \\ \\ pq_{21}E(WU_1^{'}) + (1-p)q_{22}E(WU_2^{'}) & E(WW_2^{'}) \end{bmatrix}$$

$$= \begin{bmatrix} (pq_{21} + (1-p)q_{22})I_{d\times d} & 0 \\ \\ 0 & E(WW_2^{'}) \end{bmatrix},$$

where we denote $p_1 = p$ and $p_2 = 1 - p$ in case of two classes.

To derive the CCA projection $A_2 = A_2(X, Z_1)$, the two $m \times d$ orthonormal matrices A_2 and B_2 shall maximize the singular values of $A_2^{'}\Sigma_{XZ_1}B_2$ (we take $B_2 = [b_1, \ldots, b_d]$ as in Equation (3.1), similarly to how we define A_2) [60]. Because A^* represents the dimension d subspace spanned by the first d coordinate axes, $A_2(X, Z_1)$ is optimal if and only if A_2 consists of the first d left singular vectors of Σ_{XZ_1}. Due to the form of Σ_{XZ_1}, in this case B_2 must consist of the first d right singular vectors and the respective correlations are maximized to the decreasingly ordered singular values of the $d \times d$ leading principal sub-matrix of Σ_{XZ_1}. Therefore $A_2A_2^{'} = A^*A^{*'}$ if and only if A_2 is spanned by the first d coordinate axes, or equivalently the largest d singular values of Σ_{XZ_1} all come from the $d \times d$ leading principal sub-matrix.

Putting into inequalities, the CCA projections $A_2(X, Z_s)$ are optimal if and only if

$$h_s = pq_{s1} + (1-p)q_{s2} - \sigma_1(E(WW_s^{'})) > 0. \tag{3.5}$$

When either CCA projections is not optimal, at least one h_s is non-positive and represents the "singular value loss" of using CCA.

To derive the GCCA projection A_3 based on (X, Z_1, Z_2), the covariance matrix

between Z_1 and Z_2 also comes into play:

$$\Sigma_{Z_1Z_2} = \begin{bmatrix} pq_{11}q_{21}E(U_1U_1') + (1-p)q_{12}q_{22}E(U_2U_2') & pq_{11}E(U_1W_2') + (1-p)q_{12}E(U_2W_2') \\ \\ pq_{21}E(W_1U_1') + (1-p)q_{22}E(W_1U_2') & E(W_1W_2') \end{bmatrix}$$

$$= \begin{bmatrix} (pq_{11}q_{21} + (1-p)q_{12}q_{22})I_{d\times d} & 0 \\ \\ 0 & E(W_1W_2') \end{bmatrix}.$$

Argued in a similar manner, the GCCA projection is optimal if and only if A_3 is spanned by the first d coordinate axes. The necessary and sufficient condition for that is

$$h + h_1 + h_2 > 0, \tag{3.6}$$

where we define $h = pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(W_1W_2'))$. In words, if both the CCA projections are already optimal, it is sufficient that the largest d singular values of $\Sigma_{Z_1Z_2}$ all come from the $d \times d$ leading principal sub-matrix; else if either CCA projections is not optimal, the "singular value gain" from $\Sigma_{Z_1Z_2}$ has to compensate the possible "singular value loss" from Σ_{XZ_1} and Σ_{XZ_2} in order for the GCCA projection to be optimal.

An important step is to prove that if $h_s \geq 0$ for $s = 1, 2$, then $h > 0$. This is true

because

$$\begin{aligned} h &= pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(W_1W_2^{'})) \\ &\geq pq_{11}q_{21} + (1-p)q_{12}q_{22} - \sigma_1(E(WW_1^{'}))\sigma_1(E(WW_2^{'})) \\ &\geq pq_{11}q_{21} + (1-p)q_{12}q_{22} - (pq_{11} + (1-p)q_{12})(pq_{21} + (1-p)q_{22}) \\ &= p(1-p)(q_{11} - q_{12})(q_{21} - q_{22}) \\ &> 0, \end{aligned}$$

where the first inequality uses condition (3) in Definition 2, the second inequality is by the fact that $h_s \geq 0$, and the last inequality uses condition (4).

By the above derivation, if both CCA projections are optimal such that $h_s > 0$ for $s = 1, 2$, then Inequality (3.6) automatically holds and the GCCA projection A_3 is also optimal. This shows that any $F_3 \in \Omega_3^*$ satisfying Inequality (3.5) for $s = 1, 2$ is an element of the subset $\{F_3 \in \Omega_3^* | \max\{L_{A_2}\} = L_{A_3} = L_{A^*}\}$.

Next we show there exists $F_3 \in \Omega_3^*$ such that Inequality (3.6) holds while Inequality (3.5) fails for at least one s. The trivial example is that: if $h_1 = h_2 = 0$, then the GCCA projection is optimal. Furthermore, fixing h, p and all the q_{sk}, the left-hand side of Inequality (3.6) is clearly continuous with respect to $\sigma_1(E(WW_s^{'}))$ for each s. This means $\sigma_1(E(WW_s^{'}))$ can be increased such that $h_s < 0$ (and condition (3) in Definition 2 will not be violated) while Inequality (3.6) still holds. So there also exists F_3 such that the GCCA projection is optimal when $h_s < 0$. Thus $\exists F_3 \in \{F_3 \in \Omega_3^* | \max\{L_{A_2}\} > L_{A_3} = L_{A^*}\}$.

Therefore, when A^* is unique and $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices, we proved that: for any given $F_3 \in \Omega_3^*$, if the CCA projections are optimal, so are the GCCA projections; if the CCA projections are not optimal (Inequality (3.5) is not satisfied for at least one s), the GCCA projection may be optimal (depending on whether the covariance structure satisfies Inequality (3.6)). Equivalently, we demonstrate that the similarity definition is sufficient for GCCA to improve CCA. Note that the step that ensures $h > 0$ when $h_s \geq 0$ will be used again.

Next we show that the result so far is invariant under any $H_X, H_{Z_s}, \Sigma_{Z_s}$ that satisfy Definition 2. Take CCA on (X, Z_1) for an example: by Equation (3.3) and Equation (3.4) we have $\Sigma_{\tilde{X}} = H_X \Sigma_X H_X' = I$ and $\Sigma_{\tilde{Z}_1} = H_{Z_1} \Sigma_{Z_1} H_{Z_1}'$; also by eigen-decomposition there exists $m_1 \times m_1$ matrix V s.t. $\Sigma_{\tilde{Z}_1} = V'V$. Then $\Sigma_X = H_X^{-1} H_X^{-1'}$ and $\Sigma_{Z_1} = H_{Z_1}^{-1} V' (H_{Z_1}^{-1} V')'$, and the CCA formulation (3.1) is equivalent to

$$\rho_{\{a_i' X, b_i' Z_1\}} = \frac{(H_X^{-1'} a_i)' H_X' \Sigma_{XZ_1} H_{Z_1}' V^{-1} (V H_{Z_1}^{-1'}) b_i}{\sqrt{(H_X^{-1'} a_i)' H_X^{-1'} a_i} \sqrt{(V H_{Z_1}^{-1'} b_i)' V H_{Z_1}^{-1'} b_i}},$$

$$\text{subject to } \rho_{\{a_i' X, a_j' X\}} = \frac{(H_X^{-1'} a_i)' H_X^{-1'} a_j}{\sqrt{(H_X^{-1'} a_i)' H_X^{-1'} a_i} \sqrt{(H_X^{-1'} a_j)' H_X^{-1'} a_j}} = 0$$

$$\text{and } \rho_{\{b_i' Z_1, b_j' Z_1\}} = \frac{(V H_{Z_1}^{-1'} b_i)' V H_{Z_1}^{-1'} b_j}{\sqrt{(V H_{Z_1}^{-1'} b_i)' V H_{Z_1}^{-1'} b_i} \sqrt{(V H_{Z_1}^{-1'} b_j)' V H_{Z_1}^{-1'} b_j}} = 0,$$

where V^{-1} is defined as the unique Moore-Penrose pseudo inverse if $\Sigma_{\tilde{Z}_1}$ is singular. Hence it is equivalent to consider the projections $H_X^{-1'} A_2$ and $V H_{Z_1}^{-1'} B_2$ on $(\tilde{X}, V^{-1'} \tilde{Z}_1)$ (both $\tilde{X}$ and $V^{-1'} \tilde{Z}_1$ are of identity variance) with covariance

$$H_X' \Sigma_{XZ_1} H_{Z_1}' V^{-1},$$

instead of the projections A_2 and B_2 on (X, Z_1). The same holds for the GCCA formulation (3.2). Furthermore, the classification task remains the same because the projected feature $A^{'}X = (H_X^{-1'}A)^{'}H_X X$ is invariant under the full-rank transformation H_X. Therefore the optimal projection A^* and the GCCA/CCA projections A_{s+1} are all equivalent to the identity variance case up to H_X, and the result is clearly invariant.

At last we justify the case when A^* is not unique, which means there exists A^* that is spanned by the first d coordinate axes under different transformation matrices. Because the conditions in Definition 2 are required to be satisfied for all A^*, in most cases the CCA optimality is still equivalent to Inequality (3.5), i.e., CCA is optimal if and only if Inequality (3.5) is satisfied for at least one A^* after proper transformations for each A^*. The same holds for the GCCA optimality (Inequality (3.6)), and we can still conclude that GCCA improves CCA following the same steps. However, a special case should be taken into consideration, and we take the CCA projection $A_2(X, Z_1)$ for an illustration: Suppose the singular vector $\sigma_1(E(WW_s^{'}))$ corresponds to is the $(d+1)$th coordinate axes and $\sigma_1(E(WW_s^{'})) > \sigma_2(E(WW_s^{'}))$. Then $A_2(X, Z_1)$ can be chosen to represent any dimension d subspace of the space spanned by the first $(d+1)$ coordinate axes, and the degrees of freedom is $(d+1)d - \frac{d^2+d}{2}$ (the degrees of freedom may increase if there are repeating singular values). Now, if A^* happens to have the same degrees of freedom in the space spanned by the first $(d+1)$ coordinate axes, then $A_2(X, Z_1)$ is optimal if and only if $h_1 \geq 0$ (rather than $h_1 > 0$) because

any arbitrary choice of A_2 is optimal. Similar phenomenon applies for A_{s+1}, in which case Inequality (3.5) and Inequality (3.6) should be adjusted to include equalities. However, in this case we still have $h + h_1 + h_2 > 0$ when the CCA projections are optimal, which is still sufficient (but may not be necessary) for GCCA to be optimal. Therefore, GCCA still improves CCA in case of non-unique A^*, and the justification is done. □

3.7.2 Proof of Theorem 7 for any $K \geq 2$ and $r \geq 1$

Proof. Now we generalize the result to arbitrary $K \geq 2$ (multi-class) and any $r \geq 1$ (the GCCA criterion). Without loss of generality, we assume that A^* is unique and $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices.

Let us treat the case that $r = 1$ first. Using the setting in Equation (3.4) and argue similarly as before, GCCA improves CCA if and only if

$$h = \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W_1 W_2^{'})) > 0 \tag{3.7}$$

is true when $h_s = \sum_{k=1}^{K} p_k q_{sk} - \sigma_1(E(W W_s^{'})) \geq 0$ for $s = 1, 2$.

This is true because

$$\begin{aligned} h &= \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W_1 W_2^{'})) \\ &\geq \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W W_1^{'}))\sigma_1(E(W W_2^{'})) \\ &\geq \sum_{k=1}^{K} p_k q_{1k} q_{2k} - (\sum_{k=1}^{K} p_k q_{1k})(\sum_{k=1}^{K} p_k q_{2k}) \\ &= \sum_{1 \leq k_1 < k_2 \leq K} p_{k_1} p_{k_2} (q_{1k_1} - q_{1k_2})(q_{2k_1} - q_{2k_2}) \qquad (3.8) \\ &> 0, \end{aligned}$$

where the first inequality follows from conditions (3), the second inequality follows from $h_s \geq 0$, the next equality follows from simple algebra, and the last inequality follows from condition (4).

As to the GCCA criterion with $r \geq 1$, GCCA improves CCA if and only if

$$(\sum_{k=1}^{K} p_k q_{1k} q_{2k})^r - \sigma_1^r(E(W_1 W_2^{'})) > 0$$

is true when $h_s \geq 0$. Clearly this inequality holds if and only if it holds for $r = 1$, which is Inequality (3.7). Hence it is true and GCCA improves CCA in the similar family for any $r \geq 1$.

Thus Theorem 7 is proved for any number of classes and any GCCA criterion with $r \geq 1$. □

3.7.3 Proof of Corollary 1 and Corollary 2

Proof. Without loss of generality, we carry out the proof assuming A^* is unique, $H_X, H_{Z_s}, \Sigma_{Z_s}$ are all identity matrices, and $K = 2$ and $r = 1$.

There are S auxiliary features in total, and thus $\binom{S}{S'}$ choices of auxiliary features for $A_{S'+1}$. We define $h_s = pq_{s1} + (1-p)q_{s2} - \sigma_1(E(WW_s^{'}))$ and $h_{st} = pq_{s1}q_{t1} + (1-p)q_{s2}q_{t2} - \sigma_1(E(W_sW_t^{'}))$ for any s and t satisfying $S \geq s, t \geq 1$, where h_{st} is a generalization of h in the proof of Theorem 7.

Then the GCCA projection $A_{S'+1}$ using the first $S^{'}$ auxiliary features is optimal if and only if

$$\sum_{1 \leq s < t \leq S'} h_{st} + \sum_{s=1}^{S'} h_s > 0. \tag{3.9}$$

This is a generalization of Inequality (3.6), because there are $S^{'}$ possible "singular value loss" caused by Σ_{XZ_s} and $\frac{S^{'}(S^{'}-1)}{2}$ additional cross-covariance terms $\Sigma_{Z_sZ_t}$ between the auxiliary features. Note that for any other $A_{S'+1} \in \mathcal{A}_{S'+1}$ with a different choice of auxiliary features, we can still use Inequality (3.9) for the optimality by switching the first $S^{'}$ auxiliary features with the chosen $S^{'}$ auxiliary features.

All the CCA projections are optimal if and only if $h_s > 0$ for all $s = 1, \ldots, S$. This implies that $h_{st} > 0$ is always true for any $1 \leq s < t \leq S$, and Inequality (3.9) holds for any $A_{S'+1} \in \mathcal{A}_{S'+1}$ with $S \geq S^{'} \geq 2$. Therefore the set of GCCA projections $\mathcal{A}_{S'+1}$ always improves the set of CCA projections $\mathcal{A}_2$, and Corollary 1 is proved.

To prove Corollary 2, we use the simplifying condition (4'). Then Inequality (3.9) simplifies to $\frac{S^{'}-1}{2}h_{12} + h_1 > 0$, because h_{st} are the same for all $1 \leq s, t \leq S^{'}$ and so

are h_s. We need to show that if $A_{S'}$ are optimal for certain F_{S+1}, so is $A_{S'+1}$. (note that the choice of auxiliary features no longer matters because they follow the same distribution, which means all the elements in $\mathcal{A}_{S'+1}$ represent the same subspace.)

When $S' = 2$, it is a special case of Theorem 7 because any F_{S+1} satisfying condition (4') also satisfies condition (4). Clearly A_2 is optimal if and only if $h_1 = h_2 > 0$, which implies $h_{12} > 0$. So Inequality (3.9) holds and A_3 is also optimal.

When $S' = 3$, A_3 is optimal if and only if $h_{12} + h_1 > 0$. In this case if $h_1 > 0$, then we have $h_{12} > 0$; if $h_1 < 0$, then $h_{12} > 0$ must be true in order for A_3 to be optimal. In any case, $\frac{3}{2}h_{12} + h_1 > 0$ is true and A_4 is optimal.

Therefore, the optimality of A_3 implies the optimality of A_4. By induction, for any $S \geq S' \geq 2$, the optimality of $\mathcal{A}_{S'}$ implies the optimality of $A_{S'+1}$ under the simplifying condition (4'), and Corollary 2 is proved. Note that the corollary is not true under the original condition (4), and one can easily make up a counter-example by checking Inequality (3.9). □

3.7.4 Comments

We conclude the proof section by considering the term $h = \sum_{k=1}^{K} p_k q_{1k} q_{2k} - \sigma_1(E(W_1 W_2'))$ in Equation 3.7 for the case of two auxiliary features, which offers additional insights for Definition 2 of the similar family and is potentially useful for model selection.

Firstly, the equation offers a relaxation of condition (4) in the similar family:

instead of $(q_{sk_1} - q_{sk_2})(q_{tk_1} - q_{tk_2}) > 0$ for all $1 \leq s < t \leq S$ and $k_1, k_2 = 1, \ldots, K$, we can replace it by either $h > 0$ or $\sum_{1 \leq k_1 < k_2 \leq K} p_{k_1} p_{k_2} (q_{1k_1} - q_{1k_2})(q_{2k_1} - q_{2k_2}) > 0$ (by Equation 3.8), which is more difficult to interpret than the original condition but less restrictive.

Secondly, the improvement of GCCA over CCA depends almost solely on the magnitude of h. The larger the h, the more likely that GCCA may be optimal even if CCA is not. Towards this direction, the magnitude of q_{sk} plays an important role: for fixed $E(W_1 W_2')$, assuming all coefficients non-negative, h increases with q_{sk} and GCCA projection is potentially more superior.

Finally, the above observation may be useful for the choice of auxiliary variables and the projecting dimension without using cross-validation. Other things being equal, an auxiliary variable with larger h or q_{sk} is more favorable, as is a projection dimension with larger h or q_{sk}; thus it is reasonable to choose an auxiliary variable and/or a projection dimension with a more significant "signal" part (where U_k lives) for later inference, which agrees with intuition. Numerically, within the similar family this observation is useful for model selection purposes (choose the auxiliary feature and/or the projection dimension with the largest h using greedy algorithms, among all available auxiliary features and all possible dimensions); but out of the similar family definition, whether a modified version of h can serve the model selection purpose or not requires further investigation.

Chapter 4

Nonlinear Manifold Matching

4.1 Introduction

So far we have shown that separate projection can be harmful for matching in Chapter 2, and joint projection can be useful for later inference in Chapter 3. But we have limited the scope to traditional linear projections, i.e, principal component analysis (PCA) and canonical correlation analysis (CCA). And as mentioned in Chapter 1, unfolding the non-linearity can be beneficial for subsequent inference, and many manifold learning algorithms have been proposed to learn the intrinsic low-dimensional structure of nonlinear data.

In this chapter, we apply nonlinear transformations to the manifold matching task for two or more data sets from disparate sources. There are many recent endeavors in data fusion and manifold matching [61], [25], [31], [32], [19]; and similar

to dimension reduction of a single data set, manifold matching usually serves as a feature extraction step to explore multiple data sets, and has also been shown to help subsequent inference in object recognition, information retrieval and transfer learning [62], [56], [34], [35], [18], [63].

Due to the success of nonlinear embedding algorithms for a single data set, it seems intuitive that they can be combined with proper matching methods to achieve better feature extraction for multiple data sets. The simplest procedure is to pick one nonlinear algorithm, apply it to each data set separately, then match the transformed data sets together. But there exists three questions for this simple procedure: Firstly, how to assess whether the nonlinear algorithm improves the matching task? Secondly, among so many nonlinear embedding algorithms, each has its pros and cons; which algorithm is most suitable for the matching task? Thirdly, can we optimize the procedure and achieve better performance for disparate data matching, comparing to the separate embed and match strategy? To that end, we use distance correlation and hypothesis testing power to evaluate the matching quality, and propose a nonlinear manifold matching algorithm using shortest-path distance and joint neighborhood selection. The algorithm turns out to significantly improve the matching quality for disparate data matching (e.g., one data set has nonlinear geometry while the other does not), and also achieves robust performance against model selection and noisy data.

This chapter is organized as follows: In Section 4.2 we review the basic setting,

three common matching methods, followed by the Isomap algorithm that constructs the shortest-path distance. In Section 4.3 we present our nonlinear manifold matching algorithm, the evaluation criteria using distance correlation and hypothesis testing, and discuss various algorithmic issues. In Section 4.4 we illustrate the advantages of our methodology via numerical simulations and real data experiments, using the simulated Swiss roll data and the Wikipedia document data with text and graph features.

Note that this chapter is based on the paper [21]; related code and data are available on the website [1].

4.2 Reviews

4.2.1 The Matching Framework

We first introduce a formal setting for multiple matched data sets, and then briefly review the three matching methods discussed in [19].

Suppose n objects are measured under two different sources. Then we have available $X_l = \{x_{il}\} \in \Xi_l$ for $l = 1, 2$, $i = 1, \ldots, n$, with $x_{i1} \sim x_{i2}$ for each i ($\sim$ denotes the matched data points). We assume $x_{il} \in \mathbb{R}^m$, or equivalently $\Xi_l = \mathbb{R}^{m \times n}$. Note that in practice matched data of disparate sources may have different or even unknown dimensions, say an image and its description, in which case it is more appropriate to

[1] `http://www.cis.jhu.edu/~cshen/`

assume that each space Ξ_l is endowed with a distance measure. But as the data can always be embedded into a proper ambient dimension first, for ease of presentation we assume the ambient space is $\mathbb{R}^m$ for all data sources. The setting is also extendable to more than two data sets, but for convenience we assume $l = 2$ for most of the chapter.

Since the ambient dimension m is usually large in modern applications, dimension reduction is often required to achieve a meaningful matching. The matching and embedding of multiple data sets are formulated as finding two mappings $\rho_l : \mathbb{R}^m \to \mathbb{R}^d, l = 1, 2$ based on the given data X_l, i.e., ρ_l embed and match the two data sets in the common low-dimensional space $\mathbb{R}^d$. Here d should satisfy $1 \leq d \leq m$, and we denote $\hat{X}_l = \{\rho_l(x_{il})\}$ as the mapped data in $\mathbb{R}^d$. There are many ways to assess the matching quality, but intuitively we would like the matched pairs $(\hat{x}_{i1}, \hat{x}_{i2})$ to be as close as possible for all i, while unmatched data are not close.

We presented three matching methods in [19] to derive ρ_l based on different objective functions, namely MDS followed by the Procrustes matching, CCA matching, and joint MDS. In what follows, X_l represents an $m \times n$ data matrix properly centered for each l, and the final output $\hat{X}_l$ is a $d \times n$ matrix.

The Procrustes method first projects the data separately by MDS or PCA into $\mathbb{R}^d$, then minimizes the Procrustes fit $\|PU_1X_1 - U_2X_2\|_F^2$ by finding a proper rotation P. Here U_l is the $d \times m$ PCA projection for each X_l, P is the $d \times d$ Procrustes transformation, and $\|\cdot\|_F$ is the Frobenius norm. Thus the two mappings are $\rho_1 = PU_1$

and $\rho_2 = U_2$.

The CCA method finds two $d \times m$ CCA transformations C_l to maximize the correlation between $\hat{X}_1 = C_1 X_1$ and $\hat{X}_2 = C_2 X_2$, subject to the constraints that the sample covariance matrix of $C_l X_l$ is identity for each l. Thus the two mappings are $\rho_1 = C_1$ and $\rho_2 = C_2$.

The joint MDS method constructs a $2n \times 2n$ distance matrix using Euclidean distance within each X_l, and then applies MDS to directly project the data into $\mathbb{R}^d$. Note that the off-diagonal distance, i.e., the distance between X_1 and X_2, is usually unavailable and needs to be properly imputed; details are in Section 4.3.

We have investigated the property of these matching methods [19], [17], but it is not our purpose here to discuss which one is better for matching. They are all intuitive to use and easy to implement, and serve as a platform for introducing nonlinear embedding into the matching framework, because different nonlinear embedding algorithms may work better with certain matching method than others.

4.2.2 Shortest-Path Distance and Isomap

As we use shortest-path distance for matching, Isomap is the main nonlinear algorithm that applies shortest-path distance to achieve nonlinear embedding. Thus we review the Isomap algorithm here followed by some discussions. In this context we use X_1 to denote the original data, $\hat{X}_1$ to denote the embedded data by Isomap, and d as the embedding dimension.

The algorithm works as follows: First it constructs a nearest-neighbor graph G based on the Euclidean distance matrix $\Delta_1(ij) = \|x_i - x_j\|$, by k-nearest-neighbor (kNN) method or ϵ-ball method. Then this graph is used to iteratively calculate the shortest-path distance matrix Δ_G. Finally $\hat{X}_1$ is obtained by projecting Δ_G into dimension d by MDS.

The other steps in the Isomap algorithm being routine, the shortest-path distance construction, i.e., the calculation of Δ_G, is the key step: For each i, j, initiate $\Delta_G(i,j) = \Delta_1(i,j)$ if x_{i1} and x_{j1} are adjacent in G, $\Delta_G(i,j) = \infty$ otherwise. Then iterate through $q = 1, \ldots, n$ and replace the entry $\Delta_G(i,j)$ by $\min\{\Delta_G(i,j), \Delta_G(i,q) + \Delta_G(q,j)\}$. The final matrix Δ_G becomes the shortest-path distance matrix for X_1. This can be effectively implemented by Floyd's algorithm or Dijkstra's algorithm.

Therefore, Isomap is essentially the same as MDS except it constructs the shortest-path distance matrix for MDS application rather than the original distance matrix. It has been shown in [64], [9] that the shortest-path distance can recover the geodesic distance of isometric manifolds with high probability under certain sampling condition, and can also recover the geodesic distance of certain curved manifolds when using a slightly different version called conformal Isomap [65].

Other than the fact that Isomap cannot recover all types of nonlinear geometry, the main downside of Isomap is the running time for large n. But its computation can be sped up by landmark Isomap or out-of-sample MDS, see [9], [66], [67]. The idea is to pick a subset of landmark points to do usual Isomap, then compute the shortest-

path distance of all other points with respect to the landmark points only, followed by out-of-sample MDS embedding. In this chapter we do not use this technique for speed purpose, because we do not use large n in the experiments. But the out-of-sample technique is applied in hypothesis testing as described in the next section.

4.3 Manifold Matching Framework

In this section we first present the nonlinear manifold matching algorithm using shortest-path distance and joint neighborhood selection, then propose two evaluation criteria (distance correlation and hypothesis testing), followed by discussions on various implementation issues.

4.3.1 Main Algorithm

Our algorithm can be decomposed into three steps, where the first step applies joint neighborhood selection, the second step constructs the shortest-path distance, and the last step embeds and matches the data based on the constructed distances.

Step 1: Jointly select the neighbors and construct a single nearest-neighbor graph G for all data $\{X_l, l = 1, 2\}$. This can be achieved by using the sum of distance to derive the nearest neighbors, i.e., whether x_{il} is adjacent to x_{jl} in G is determined by $\sum_l \Delta_l(i, j)$ instead of $\Delta_l(i, j)$, and we always use k-nearest-neighbor for neighborhood selection.

Note that in order to achieve a meaningful joint neighborhood selection, it is necessary to pre-scale the data so the distance matrices are on the same scale. Alternatively, one may use a weighted sum of distance or rank-based method to derive a joint neighborhood.

Step 2: Given the graph G, calculate the shortest-path distance matrices Δ_{G_l} for each l using the same procedure as Isomap.

Step 3: Derive the mappings ρ_l and low-dimensional mapped data $\hat{X}_l$ using any of the three matching methods on $\{\Delta_{G_l}\}$.

Specifically, for the Procrustes method, we separately embed the two shortest-path distance matrices into $\mathbb{R}^d$ by MDS, followed by Procrustes transformation.

For the CCA method, the same separate MDS embeddings are matched by CCA transformations in $\mathbb{R}^d$ to maximize the correlation.

For the joint MDS method, we concatenate an omnibus matrix

$$M_G = \begin{bmatrix} \Delta_{G_1} & O_G \\ \\ O_G' & \Delta_{G_2} \end{bmatrix} \tag{4.1}$$

with $O_G = (\Delta_{G_1} + \Delta_{G_1})/2$, and apply MDS (either classical MDS or raw-stress with proper weights, see [19]) directly on M_G to yield the embeddings $\{\hat{X}_l, l = 1, 2\}$ in $\mathbb{R}^d$.

Note that if we only use step 3 of the algorithm without step 1 and step 2, it is equivalent to match the original distance without any nonlinear algorithm.

4.3.2 Evaluation Criteria

To assess the quality of the nonlinear manifold matching algorithm, we use distance correlation and hypothesis testing power as the evaluation criteria.

The distance correlation proposed in [20], [68] measures the correlation between data by distance, and has the nice property of being 0 if and only if two random variables are independent. The notion of distance correlation is particularly suitable for evaluating the shortest-path distance constructed in step 2 of our algorithm. If the distance correlation between the jointly constructed shortest-path distances is significantly larger than others (such as the distance correlation between the original distances, or between the shortest-path distances without joint neighborhood, or between the Euclidean distances by other algorithms like LLE), it indicates that shortest-path distance and joint neighborhood are better for matching.

We also consider the following hypothesis test used in [19]. First we randomly split the sample data into training data X_l of matched pairs, testing data $Y_l = \{y_{il}\}$ containing both matched pairs and unmatched pairs, and consider the matching test $H_0 : y_{i1} \sim y_{i2}$. Then we learn the mappings ρ_l based on the training data X_l only, apply the mappings to the testing data, and use the Euclidean distance $T = d(\hat{y}_{i1}, \hat{y}_{i2})$ in the embedded space as the test statistic. We can construct the power curve by calculating the empirical distributions of the test statistic T under the null and the alternative, for which a higher matching power indicates a better manifold matching algorithm.

Those two criteria are complementary: distance correlation is fast to compute and independent of the embedding dimension and the matching method, but it is possible that the improvement of distance correlation can be offset by proper matching or may be due to over-fitting; hypothesis testing is a more complete evaluation of the manifold matching algorithm, but the testing power depends on the dimension choice in step 3. Thus we use both criteria in this chapter. Note that there exists many other possible criteria: for example, one may test the usual correlation or the Procrustes statistic on the embedded data; and if the sample data has labels, one may test the classification error after matching, etc.

4.3.3 Discussions

In this subsection we discuss some implementation details and potential extensions of the nonlinear manifold matching algorithm, as well as providing explanations for using joint neighborhood selection and shortest-path distance.

On the scaling and centering of the data: To obtain a meaningful matching, proper scaling is usually required. This can be achieved by scaling the original data X_l, or the shortest-path distance matrices Δ_{G_l}, or the embedded data $\hat{X}_l$. And many algorithms such as conformal Isomap, LLE, and LTSA do implicit scaling in their algorithms. In order to facilitate the joint neighborhood step, we always pre-scale the original data to have the same Frobenius norm.

In addition to scaling, centering the original data X_l or the embedded data $\hat{X}_l$ to

have the same mean is also necessary for matching, which is implicitly handled by all embedding algorithms used in the chapter.

On the joint neighborhood selection: Comparing to the usual separate neighborhood selection, it is more intuitive to consider joint neighborhood selection for a better manifold matching: in the ideal matching case, if x_{i1} is adjacent to x_{j1} in the first data set, so should x_{i2} and x_{j2} in the second data set; and in case of noisy data, separate graphs may yield larger discrepancy for later nonlinear embedding. Thus the two shortest-path distances should be more similar to each other when using joint neighborhood. Note that for hypothesis testing, we do not use joint neighborhood for the testing data because the testing pairs may be unmatched.

On the neighborhood size and dimension choice: The model selection problem is important for any algorithm involving nearest-neighbor graph or dimension reduction, which is also intrinsic to our manifold matching task. It has been argued that the neighborhood size k should neither be too small nor too large in order to recover the local geometry for nonlinear embedding; and there exists extensive discussions and adaptive methods on choosing the neighborhood size in [11], [69], [16]. As for the dimension choice d, it affects the embedding step and later inference based on the embedding; and there exists automatic procedures using profile likelihood or Bayesian model selection from [70], [57], [58]. Note that k is required to be larger than d in most nonlinear embedding algorithms except Isomap.

In the numerical experiments we simply choose k as 10 or 20, and we choose the

embedding dimension d based on the scree plot of the data. Although we do not delve into the complexity of the model selection problem, we provide a numerical example in Figure 4.7 to show that our manifold matching algorithm is fairly robust against the choice of k and d, due to the nature of the matching task and the evaluation criteria.

On the out-of-sample technique for testing: As already mentioned in Isomap, during the hypothesis testing of the proposed manifold matching algorithm, we keep the training data X_l as the landmark points, and use the out-of-sample technique to embed the testing data. The reason is similar to why we do not use joint neighborhood in testing: the training pairs are always matched while the testing pairs may be unmatched, and we do not wish the unmatched data to affect the distance calculation of the matched training data or the other way around. Note that the out-of-sample technique is widely available to many nonlinear algorithms like LLE, Laplacian eigenmaps, kernel PCA, see in [71], [66], [72], [73]. But we do not apply out-of-sample embedding other than using Isomap or joint neighborhood, because it does not help the testing power for non-distance based algorithms and separate neighborhood.

On using other embedding algorithms: The manifold matching algorithm can be extended to most other nonlinear embedding algorithms not limited to shortest-path distance and Isomap. Taking LLE as the example, we may keep the joint neighborhood in step 1, and use LLE in step 2 instead of the shortest-path distance. This

means the output of step 2 is two Euclidean data sets of dimension d rather than two distance matrices, and the matching step can be applied directly, i.e., in case of Procrustes or CCA matching we can directly match the data without doing separate MDS, and in case of joint MDS we form two Euclidean distance matrices to concatenate the omnibus matrix. In the numerical section, we will use LLE, LTSA, and Laplacian eigenmaps to compare with our proposed manifold matching algorithm using shortest-path distance and joint neighborhood.

On using the shortest-path distance: No nonlinear algorithm can always recover the nonlinear geometry; and shortest-path distance does not always approximate the geodesic distance either. But the matching task and the evaluation criteria ask that the matched data are close to each other, rather than that the geometry is fully recovered: the shortest-path distance is always no smaller than the original distance, and is able to enlarge the distance that is not in the local neighborhood. This often improves the distance correlation and the testing power.

Furthermore, Isomap is usually able to preserve the local geometry more faithfully than others. Many other nonlinear algorithms involve a normalization step, which only preserves the local geometry up to some affine transformation [74]. This may cause trouble in matching disparate data if the affine transformations of each data set are significantly different. Indeed, in the numerical section we will observe that normalization-based algorithms like LLE, LTSA, and Laplacian eigenmaps may not perform stably for matching noisy data or data from disparate sources, and shortest-

path distance usually prevails. But it does not mean that other nonlinear algorithms should not be used; they may still be valuable for certain data type or other evaluation criteria.

On matching more than two data sets: The manifold matching algorithm is readily extendable to match more than two data sets, which will appear in the numerical section. In this case, other than the minimal change of joint neighborhood in step 1, our algorithm needs proper modifications for the matching part in step 3: for the Procrustes method, we consider minimizing the square sum of Procrustes fit $\|Q_1\hat{X}_1 - \hat{X}_3\|_F^2 + \|Q_2\hat{X}_2 - \hat{X}_3\|_F^2 + \|Q_1\hat{X}_1 - Q_2\hat{X}_3\|_F^2$ using two Procrustes transformation matrices; for the CCA method, generalized CCA [39], [42], [18] is the standard extension, which finds three transformations C_l to maximize the sum of pair-wise correlations subject to proper constraints; for the joint MDS method, the omnibus matrix can be constructed by three distance matrices followed by proper imputation and MDS. As to the evaluation criteria, we consider the test statistic $T = d(\hat{y}_{i1}, \hat{y}_{i2}) + d(\hat{y}_{i1}, \hat{y}_{i3}) + d(\hat{y}_{i2}, \hat{y}_{i3})$ for the hypothesis test $H_0 : y_{i1} \sim y_{i2} \sim y_{i3}$, and the distance correlation is changed to the distance correlation sum (i.e., the sum of all pairwise distance correlation). The above extensions can be generalized to match any number of data sets.

4.4 Numerical Experiments

In this section we demonstrate the numerical advantages of our nonlinear manifold matching algorithm. Overall, we observe that the proposed algorithm can improve both the distance correlation and testing power, comparing to without using shortest-path distance or without using joint neighborhood.

4.4.1 Swiss Roll Simulation

The Swiss roll data from [8] is a 3D data set representing a nonlinear manifold, but intrinsically generated by points on a 2D linear manifold. Figure 4.1 shows the 3D Swiss roll data with 5000 points in colors, along with the embedded 2D data by MDS, Isomap and LLE at neighborhood size $k = 10$. Clearly MDS fails to recognize the nonlinear geometry while both Isomap and LLE succeed. The Isomap embedding looks similar to the original 2D linear manifold, but the geometry recovered by LLE is different from both Isomap and the original 2D data.

We carry out the first matching task as follows: for each Monte-Carlo run, we randomly pick $n = 1000$ training points from the original 2D linear manifold as the first data set X_1, and take the corresponding points on the 3D Swiss roll as the second data set X_2. Thus X_1 and X_2 are matched training data with distinct geometry, on which we can carry out our manifold matching algorithm and calculate the distance correlation. The parameters are set at $k = 10$, $d = 2$, and we use 100 matched testing

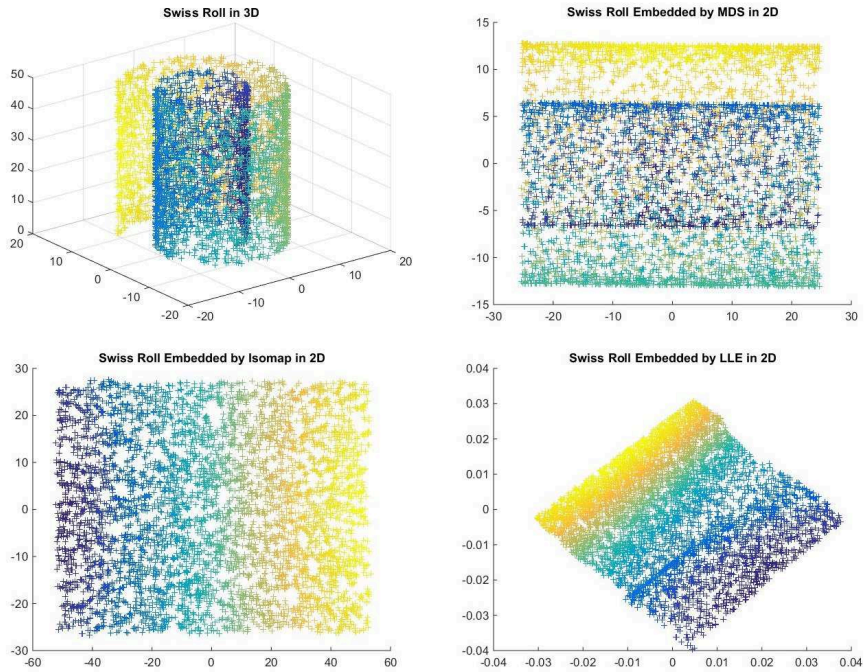


Figure 4.1: The Swiss roll data set in 3D (left top), and its 2D embedded data by MDS (right top), Isomap (left bottom) and LLE (right bottom)

pairs and 100 unmatched testing pairs to do testing.

After repeating 100 times, we present the mean distance correlation for the training data in Table 4.1, which shows that jointly constructed shortest-path distance has the best distance correlation. This advantage is reflected in the mean testing power with respect to different type 1 error levels in Figure 4.2, for which we present the matching performance for the CCA matching method combined with various nonlinear algorithms. We do not show joint MDS matching and Procrustes matching here, because their power curves have the same interpretation. We should point out that we purposely set the sample size to be 1000, because in this example the testing power for all nonlinear algorithms will converge to 1 as n increases.

Note that in the captions of all following tables and figures, original distance means that we apply the matching methods without any nonlinear embedding, joint Isomap means that we apply our proposed manifold matching algorithm using shortest-path distance and joint neighborhood selection, separate Isomap means that we separately embed the data by Isomap and then do matching, joint LLE means that we apply our algorithm using LLE and joint neighborhood selection, and separate LLE means that we separately embed the data by LLE and then do matching. Moreover, the LLE version is implemented based on the distance version presented in [11] and uses out-of-sample technique for testing, in order to compare more fairly with our proposed manifold matching algorithm. For benchmark purpose, LTSA and Laplacian eigenmaps are also added, which are not distance based and only use separate neigh-

borhood.

Table 4.1: Swiss Roll: Mean Distance Correlation for Training Data

Data Combination	Original Distance	Joint Isomap	Separate Isomap	Joint LLE	Separate LLE
(2D Linear Manifold, 3D Swiss Roll)	0.6361	**1.0000**	0.8860	0.9996	0.9394

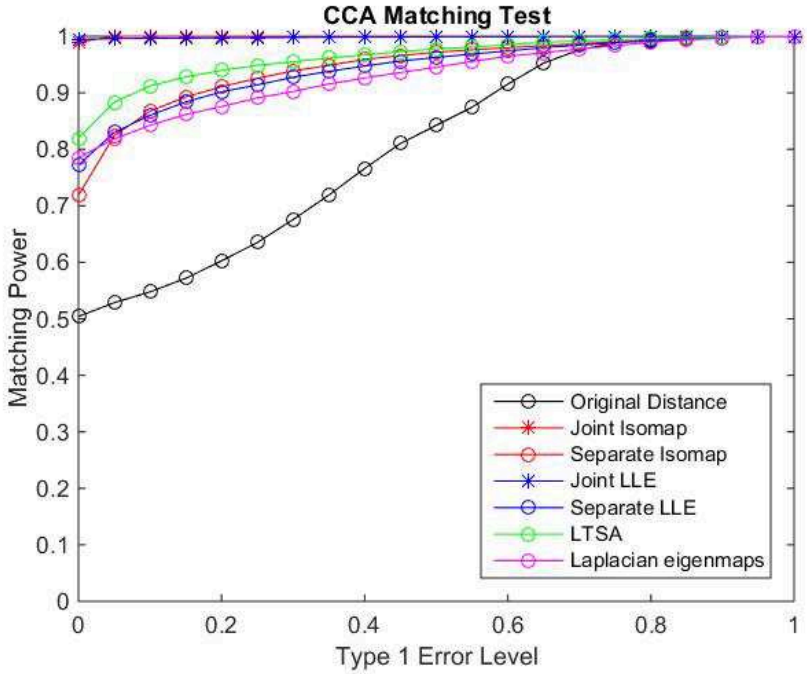


Figure 4.2: Matching Power of 3D Swiss Roll and its 2D Linear Manifold using CCA

Next we check the robustness of the manifold matching algorithm against noise. We do so by adding white noise to the first data set X_1, with the noise being independently identically distributed as $N(0, \epsilon I_{2\times 2})$. We carry out the exact same procedure as before, and plot the mean CCA matching power at the fixed type 1 error level 0.05 with respect to increasing noise $\epsilon = 0, 1, 2, \ldots, 10$ in Figure 4.3. Clearly joint Isomap is almost always superior; and even though joint LLE performs optimally in

the previous example, its sensitivity to noise degrades the matching performance for noisy data; and both Laplacian eigenmaps and LTSA are inferior to joint Isomap until $\epsilon = 9, 10$, but they are better than LLE and separate Isomap.

At last we check the performance of our algorithm for matching all linear data. We still use the original 2D linear manifold as the first data set X_1, but replace X_2 by the LLE-embedded 2D data set. Thus both data are linear with some differences, and we plot the mean CCA matching power with respect to different type 1 error levels in Figure 4.4. In this case joint Isomap performs a little better than separate Isomap, which almost coincides with matching the original distance and LTSA; and LLE and Laplacian eigenmaps performs significantly worse. This indicates that shortest-path distance and joint neighborhood are robust in matching data of similar geometry.

4.4.2 Wikipedia Articles Experiment

In this experiment we apply the manifold matching algorithm to match Wikipedia article features from disparate sources. The data contains $n = 1382$ pairs of articles from Wikipedia English and its corresponding French translations, within the 2-neighborhood of the English article "Algebraic Geometry". On Wikipedia, the same articles of different languages are almost never the exact translations of each other, because they are very likely written by different people and their contents may differ in many ways.

For both English articles and French articles, we construct a text feature matrix

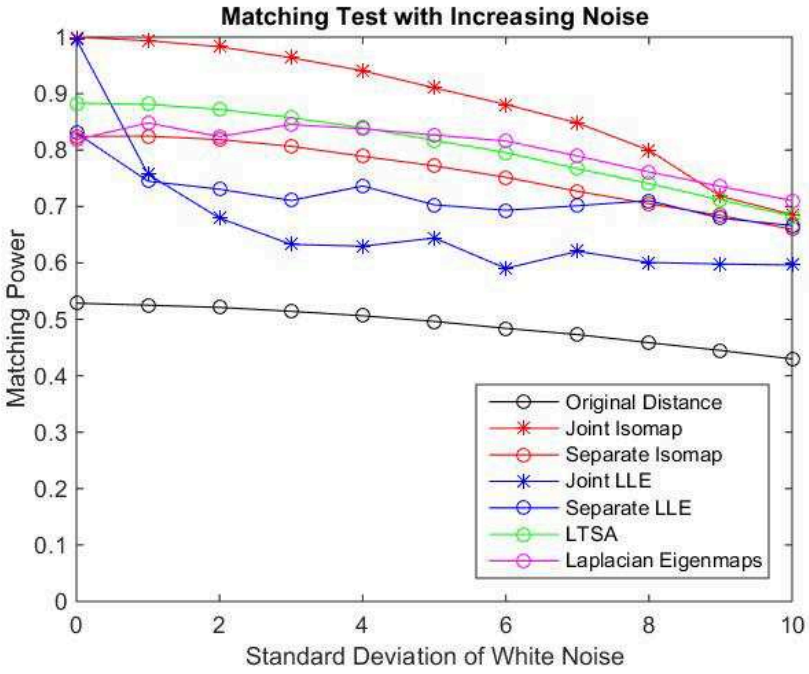


Figure 4.3: Matching Power of 3D Swiss Roll and its 2D Linear Manifold with Increasing Noise at Type 1 Error Level 0.05 using CCA

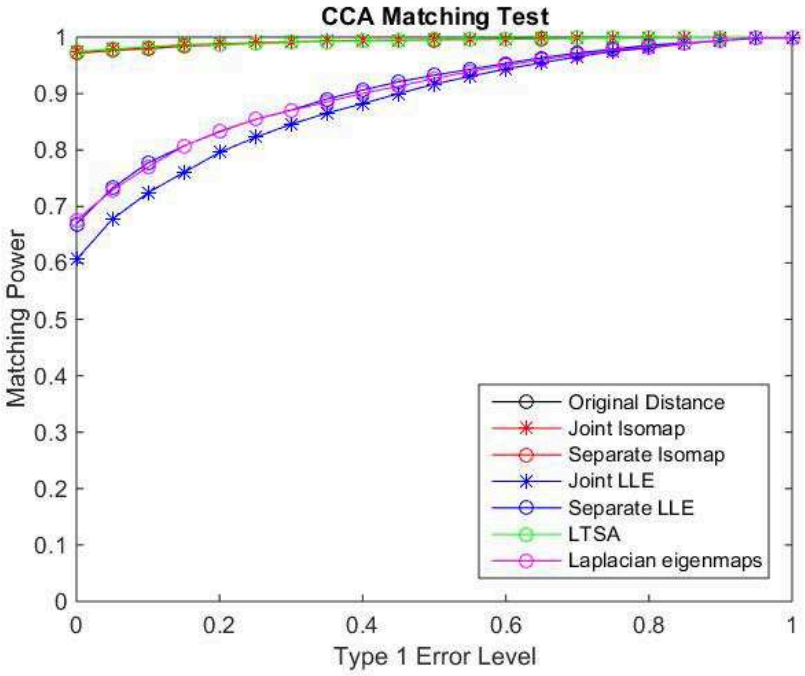


Figure 4.4: Matching Power of LLE Embedding of Swiss Roll and its 2D Linear Manifold using CCA

and a network adjacency matrix for each language: for the text feature, we consider the latent semantic indexing (LSI) features [59] followed by cosine dissimilarity to construct two text dissimilarity matrices TE and TF (standing for English text and French text); for the network, we directly construct two shortest-path distance matrices GE and GF (standing for English graph and French graph) based on the Internet hyperlinks under each language setting, and impute any path distance larger than 4 to be 6 to avoid infinite distances and scaling issues.

Thus we have four distinct matrices to describe the same article, making TE, TF, GE, GF matched in the context but of disparate sources. Furthermore, as the text matrices are derived by cosine similarity while the graph matrices are based on shortest-path distance, the former probably have nonlinear geometry while the latter should be close to linear.

As the first trial, we randomly pick $n = 500$ pairs of training points, 100 pairs of testing matched points and 100 pairs of testing unmatched points, set $k = 20$, $d = 10$, and carry out the manifold matching algorithm for every possible two data sets combination. After 100 Monte-Carlo runs, we present the mean distance correlation of the training data in Table 4.2, and the mean testing power in Table 4.3 at type 1 error level 0.05 showing only the highest matching power among Procrustes, CCA and joint MDS.

We also provide joint MDS matching power in Figure 4.5 and Figure 4.6 for matching (TE, TF) and (TE, GE) with respect to different type 1 error levels; we

do not show Procrustes and CCA matching in the figures, as they generally have the same behavior with slightly inferior matching power than joint MDS throughout the Wikipedia experiment.

Clearly joint Isomap achieves the best performance for all combinations, and LLE does not work well for either distance correlation or testing power especially when any graph matrix is involved in matching. Furthermore, from Table 4.2 and Table 4.3 we observe that our manifold matching algorithm helps the most for matching data of distinct geometry, but less significant for both linear data using graph matrices. This phenomenon is the same as the Swiss roll simulation.

As to LTSA and Laplacian eigenmaps, they cannot work with distance matrices directly. Thus for the Wikipedia data, we first project the original distance matrices into an ambient space $\mathbb{R}^m$ with $m = 50$, then proceed to apply LTSA and Laplacian eigenmaps into $\mathbb{R}^d$ followed by matching. We observe in Figure 4.5 and Figure 4.6 that they are significantly inferior to joint Isomap, with Laplacian eigenmaps performing close to the original distance matching and LTSA being much worse. This is the case for all other data combinations, so we do not show their performance in the tables. Note that it is possible that adjusting the parameter m may improve their performance, but they are not significantly better for all dimensions we tried from $m = 10$ to $m = 300$, which indicates that suitable ambient dimension may not exist for LTSA and Laplacian eigenmaps to work well for matching the Wikipedia data.

We should also point out that the actual matching power depends on the param-

eters k and d, and the power of certain algorithm can be significantly improved by changing the parameters. For example, the matching power is only 0.55 for matching (TE, TF) using original distance at $d = 10$, but it can be improved to around 0.7 at a different d; and even though joint Isomap has the best power at 0.82, it can be further increased to around 0.9 by changing d and k too. Because the model selection issue for hypothesis testing does not seem to affect the interpretation, we use the same parameters for different methods and data combinations here; also distance correlation offers an alternative performance measure independent of the embedding dimension and the matching method (except distance correlation on LLE embedding still depends on d).

Table 4.2: Wikipedia: Mean Distance Correlation for Training Data

Data Combination	Original Distance	Joint Isomap	Separate Isomap	Joint LLE	Separate LLE
(TE, TF)	0.9119	**0.9744**	0.7576	0.8626	0.8047
(TE, GE)	0.5639	**0.8039**	0.5846	0.3411	0.3139
(TF, GF)	0.5402	**0.7972**	0.5274	0.3443	0.3192
(GE, GF)	0.5740	**0.7309**	0.5708	0.3993	0.3361
(TE, GF)	0.5270	**0.8017**	0.5286	0.3472	0.3125
(TF, GE)	0.5549	**0.7944**	0.5516	0.3366	0.3134

Next we repeat the same procedure to test three data sets matching and four data sets matching. After 100 Monte-Carlo runs, we present the mean distance correlation sum in Table 4.4 and the mean testing power in Table 4.5 at type 1 error level 0.05, showing only the highest matching power among Procrustes, CCA and joint MDS.

Table 4.3: Wikipedia: Two Data Sets Matching Power at Type 1 Error Level 0.05

Data Combination	Original Distance	Joint Isomap	Separate Isomap	Joint LLE	Separate LLE
(TE, TF)	0.5505	**0.8209**	0.7851	0.4351	0.4008
(TE, GE)	0.2585	**0.5570**	0.4824	0.0967	0.0957
(TF, GF)	0.1408	**0.3227**	0.2600	0.0969	0.0998
(GE, GF)	0.2818	**0.3795**	0.3467	0.0981	0.0982
(TE, GF)	0.1506	**0.3598**	0.2927	0.0954	0.0972
(TF, GE)	0.2114	**0.4908**	0.3928	0.0956	0.0947

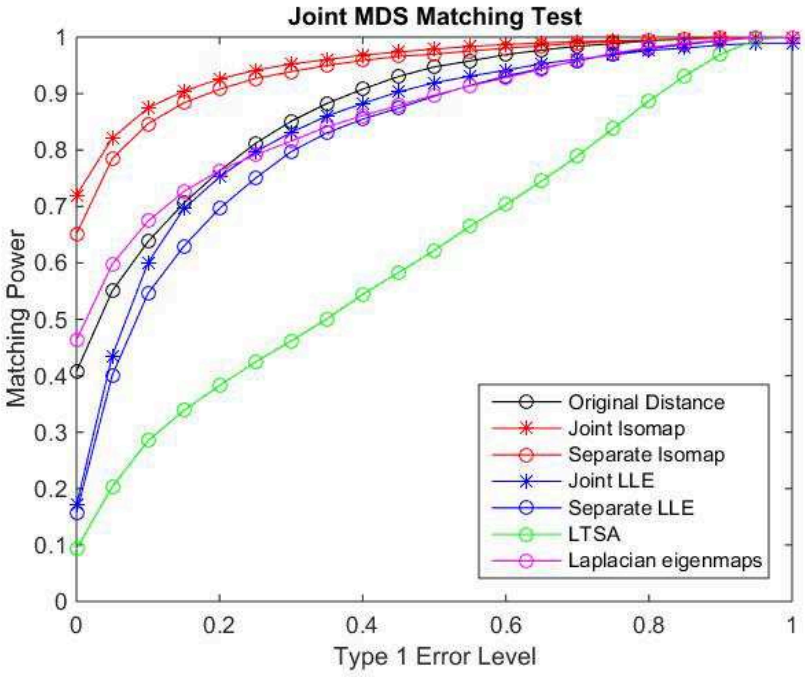


Figure 4.5: Matching Power of Wikipedia TE and TF using Joint MDS

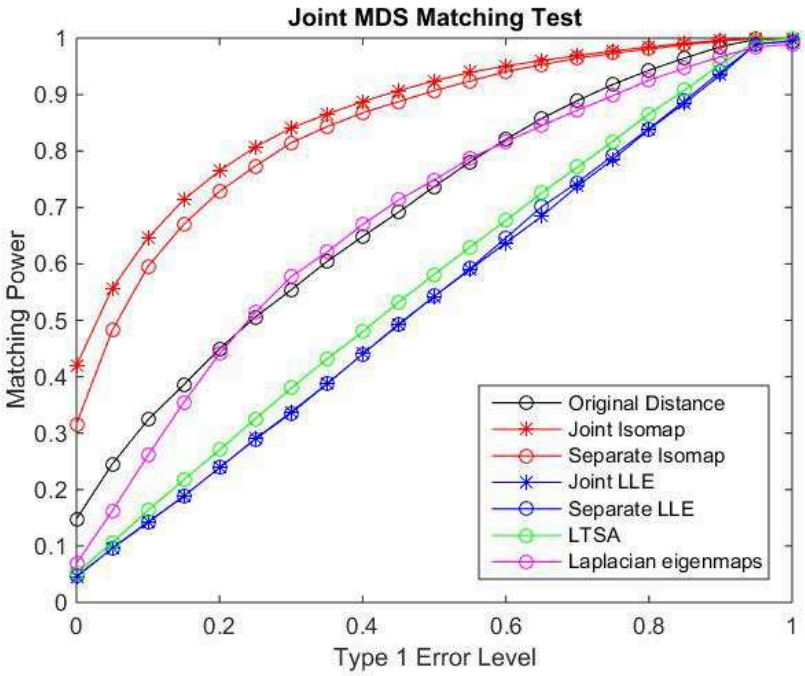


Figure 4.6: Matching Power of Wikipedia TE and GE using Joint MDS

The interpretation is similar to the two data sets matching example, which always favors our manifold matching algorithm. Note that the distance correlation sum for three data sets ranges from 0 to 3, and it ranges from 0 to 6 for four data sets matching.

We also observe that given a specific embedding algorithm, matching (TE, TF) always yields the highest distance correlation and matching power in two data sets matching, compared to other combinations of data; but adding additional graph matrix like GE and GF significantly degrades the matching power in three or four data sets matching, for all nonlinear algorithms except joint Isomap. This is probably because the network information is less reliable than the text feature; and the excel-

lent matching performance achieved by the proposed manifold matching algorithm indicates its robustness against disparate and fallible data sources.

Table 4.4: Wikipedia: Mean Distance Correlation Sum for Training Data

Data Combination	Original Distance	Joint Isomap	Separate Isomap	Joint LLE	Separate LLE
(TE, TF, GE)	2.0308	**2.5745**	1.8938	1.5691	1.4320
(TE, TF, GF)	1.9791	**2.5734**	1.8137	1.5544	1.4364
(TE, GE, GF)	1.6649	**2.3284**	1.6840	1.0791	0.9626
(TF, GE, GF)	1.6691	**2.3060**	1.6498	1.0864	0.9688
(TE, TF, GE, GF)	3.6719	**4.8843**	3.5206	2.6259	2.3998

Table 4.5: Wikipedia: More than Two Data Sets Matching Power at Type 1 Error Level 0.05

Data Combination	Original Distance	Joint Isomap	Separate Isomap	Joint LLE	Separate LLE
(TE, TF, GE)	0.3104	**0.8100**	0.7409	0.1274	0.1292
(TE, TF, GF)	0.1402	**0.6969**	0.6133	0.1353	0.1353
(TE, GE, GF)	0.1752	**0.4395**	0.3266	0.0947	0.0992
(TF, GE, GF)	0.1442	**0.4063**	0.2997	0.0952	0.0931
(TE, TF, GE, GF)	0.1539	**0.6384**	0.4309	0.1183	0.1162

At last, for the model selection issue, we show two power surface plots in Figure 4.7 for matching (TE, GE) using joint MDS matching with joint Isomap and separate Isomap respectively. The mean power is calculated at type 1 error level 0.05 with respect to different neighborhood sizes from $k = 10$ to 30 and different dimension choices from $d = 2$ to 30, for 100 Monte-Carlo runs. Note that choosing k and d for

the Swiss roll data is quite easy, because the neighborhood size has been validated to perform well and the embedding dimension equals the intrinsic true dimension (and the scree plot has a clear cut-off at $d = 2$); but for the real data, it is much more difficult to determine the optimal parameters without cross validation.

Nevertheless, Figure 4.7 shows that our approach is robust against model selection: the matching power of joint Isomap is quite stable with respect to the neighborhood size and the dimension choice, and we observe that joint neighborhood selection is consistently better than separate neighborhood selection. This phenomenon holds for Procrustes and CCA matching too, although the optimal parameters are not the same for different matching methods and different data combinations.

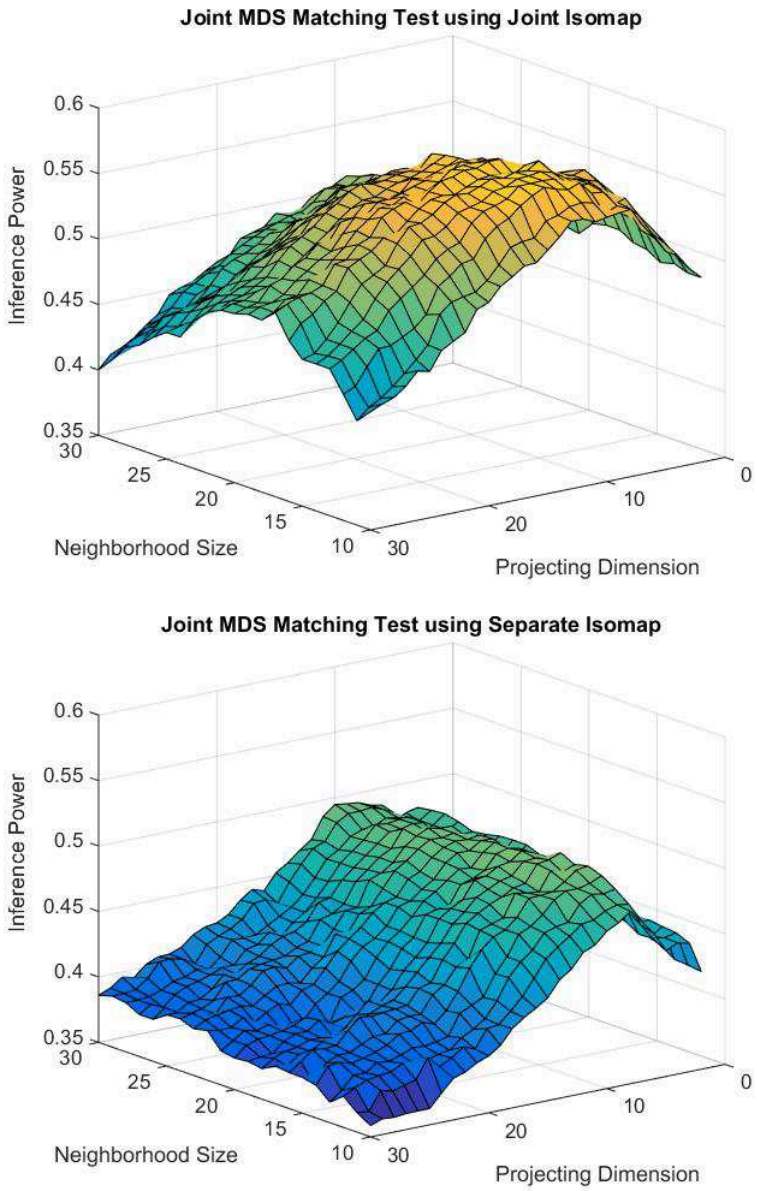


Figure 4.7: Matching Power of Wikipedia English Text and English Graph using Joint MDS with respect to Different Dimension Choices and Neighborhood Sizes at Type 1 Error Level 0.05

Chapter 5

Conclusion

The area of data analysis is a rapidly developing field, and much work is needed to fully understand how to effectively utilize the data, as the size, dimension and type of data explode in this big data age.

In this dissertation, I investigated the matching and inference performance for multiple correlated data sets. Specifically, I showed that separate projection can cause the incommensurability phenomenon in Procrustes matching; joint projection using generalized canonical correlation analysis can achieve better classification performance; nonlinear matching using shortest-path distance and joint neighborhood can increase the distance correlation and the matching quality.

Overall, I demonstrated that the inference performance for multiple correlated data sets can be significantly improved by joint matching and projection, which is illustrated by mathematical theorems and numerical experiments throughout this

dissertation.

For the future, there are many interesting research and application directions that naturally follow from the results of this dissertation. For example, how often the incommensurability phenomenon arises in practice is an important question when processing large amounts of data in a parallel computing system; the generalized canonical correlation analysis is an intuitive tool for multiple modalities, and how to easily check whether it can improve the classification error for real data is crucial to the industry; as to the nonlinear manifold matching algorithm, further improving the matching algorithm and accelerating the nonlinear transformation with theoretical guarantees are two important issues to the machine learning community.

Bibliography

[1] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.

[2] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 61, pp. 611–622, 1999.

[3] W. Torgerson, *Multidimensional Scaling: I. Theory and method.* Psychometrika, 1952.

[4] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications.* Springer-Verlag, 2005.

[5] T. Cox and M. Cox, *Multidimensional Scaling.* Chapman and Hall, 2001.

[6] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[7] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Department of Statistics, UC Berkeley, Tech. Rep., 2005.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimension reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[9] V. de Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," *Advances in Neural Informaiton Processing Systems*, vol. 15, pp. 721–728, 2003.

[10] L. K. Saul and S. T. Roweis, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[11] S. T. Roweis and L. K. Saul, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[12] D. Donoho and C. Grimes, "Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data," in *Proceedings of the National Academy of Arts and Sciences*, vol. 100, 2003, pp. 5591–5596.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[14] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[15] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimensionality reduction via tangent space alignment," *SIAM Journal on Scientific Computing*, vol. 26, no. 1, pp. 313–338, 2004.

[16] Z. Zhang, J. Wang, and H. Zha, "Adaptive manifold learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 253–265, 2012.

[17] D. E. Fishkind, C. Shen, Y. Park, and C. E. Priebe, "On the incommensurability phenomenon," *Journal of Classification*, accepted for publication.

[18] C. Shen, M. Sun, M. Tang, and C. E. Priebe, "Generalized canonical correlation analysis for classification," *Journal of Multivariate Analysis*, vol. 130, pp. 310–322, 2014.

[19] C. E. Priebe, D. J. Marchette, Z. Ma, and S. Adali, "Manifold matching: Joint optimization of fidelity and commensurability," *Brazilian Journal of Probability and Statistics*, vol. 27, no. 3, pp. 377–400, 2013.

[20] G. Szekely, M. Rizzo, and N. Bakirov, "Measuring and testing independence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[21] C. Shen and C. Priebe, "Manifold matching using shortest-path distance and joint neighborhood selection," *submitted*, http://arxiv.org/abs/1412.4098.

[22] R. Sibson, "Studies in the robustness of multidimensional scaling: Procrustes

statistics," *Journal of the Royal Statistical Society. Series B*, vol. 40, no. 2, pp. 234–238, 1978.

[23] ——, "Studies in the robustness of multidimensional scaling: Perturbation analysis of classical scaling," *Journal of the Royal Statistical Society. Series B*, vol. 41, no. 2, pp. 217–229, 1979.

[24] B. Luo and E. Hancock, "Feature matching with procrustes alignment and graph editing," in *Seventh International Conference on Image Processing And Its Applications*, vol. 1, 1999, pp. 72–76.

[25] C. Wang and S. Mahadevan, "Manifold alignment using Procrustes analysis," in *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[26] Y. Goldberg and Y. Ritov, "Local Procrustes for manifold embedding: a measure of embedding quality and embedding algorithms," *Machine learning*, vol. 7, no. 1, pp. 1–25, 2009.

[27] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford University Press, 2004.

[28] L. Qiu, Y. Zhang, and C. K. Li, "Unitarily invariant metrics on the Grassmann subspace," *SIAM Journal on Matrix Analysis and Application*, vol. 27, no. 2, pp. 507–531, 2005.

[29] T. W. Anderson, "Asymptotic theory for principal component analysis," *Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 122–148, 1963.

[30] A. W. Davis, "Asymptotic theory for principal component analysis: Non-normal case," *Australian Journal of Statistics*, vol. 19, no. 3, pp. 206–211, 1977.

[31] C. Wang, B. Liu, H. Vu, and S. Mahadevan, "Sparse manifold alignment," in *Technical Report, UMass Computer Science UM-2012-030*, 2012.

[32] A. Sharma, A. Kumar, H. D. III, and D. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[33] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[34] M. Sun and C. E. Priebe, "Efficiency investigation of manifold matching for text document classification," *Pattern Recognition Letters*, vol. 34, no. 11, pp. 1263–1269, 2013.

[35] M. Sun, C. E. Priebe, and M. Tang, "Generalized canonical correlation analysis for disparate data fusion," *Pattern Recognition Letters*, vol. 34, no. 2, pp. 194–200, 2013.

BIBLIOGRAPHY

[36] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[37] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley, 2000.

[38] J. Yang and J. Y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, vol. 36, no. 2, pp. 563–566, 2003.

[39] J. R. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.

[40] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *The Annals of Statistics*, vol. 23, no. 1, pp. 73–102, 1995.

[41] L. Sun, S. Ji, and Y. Ye, "Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 194–200, 2011.

[42] A. Tenenhaus and M. Tenenhaus, "Regularized generalized canonical correlation analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, 2011.

[43] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley Series in Probability and Statistics, 2003.

[44] Y. Chikuse, *Statistics on Special Manifolds, Lecture Notes in Statistics*. Springer, 2003.

[45] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, pp. 655–686, 2012.

[46] N. Srivastava and R. Vershynin, "Covariance estimation for distributions with 2+epsilon moments," *Annals of Probability*, vol. 41, pp. 3081–3111, 2013.

[47] D. R. Hardoon, J. Mourao-Miranda, M. Brammer, and J. Shawe-Taylor, "Unsupervised analysis of fMRI data using kernel canonical correlation," *NeuroImage*, vol. 37, no. 4, pp. 1250–1259, 2007.

[48] S. Balakrishnan, K. Puniyani, and J. Lafferty, "Sparse additive functional and kernel CCA," in *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[49] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning Journal*, vol. 83, no. 3, pp. 331–353, 2011.

[50] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[51] H. Hwang, K. Jung, and Y. Takane, "Functional multiple-set canonical correlation analysis," *Psychometrika*, vol. 77, no. 1, pp. 48–64, 2012.

[52] G. He, H.-G. Muller, and J.-L. Wang, "Functional canonical analysis for square

integrable stochastic processes," *Journal of Multivariate Analysis*, vol. 85, pp. 54–77, 2003.

[53] D. M. Witten and R. Tibshirani, "Penalized classification using Fisher's linear discriminant," *Journal of the Royal Statistical Society, Series B*, vol. 73, no. 5, pp. 753–772, 2011.

[54] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," *NIPS*, 2003.

[55] G. Fung and O. L. Mangasarian, "A feature selection newton method for support vector machine classification," *Computational Optimization and Applications*, vol. 28, no. 2, pp. 185–202, 2004.

[56] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[57] M. Zhu and A. Ghodsi, "Automatic dimensionality selection from the scree plot via the use of profile likelihood," *Computational Statistics and Data Analysis*, vol. 51, pp. 918–930, 2006.

[58] D. C. Hoyle, "Automatic PCA dimension selection for high dimensional data and small sample sizes," *Journal of Machine Learning Research*, vol. 9, pp. 2733–2759, 2008.

[59] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing

by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[60] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[61] S. Lafon, Y. Keller, and R. Coifman, "Data fusion and multi-cue data matching by difusion maps," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, 2006.

[62] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1005–1018, 2007.

[63] V. Lyzinski, D. Fishkind, and C. E. Priebe, "Seeded graph matching for correlated Erdos-Renyi graphs," *Journal of Machine Learning Research*, accepted for publication.

[64] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," 2000.

[65] V. de Silva and J. B. Tenenbaum, "Unsupervised learning of curved manifolds," *Nonlinear Estimation and Classification*, 2002.

[66] Y. Bengio, J. F. Paiement, and P. Vincent, "Out-of-sample extensions for LLE,

Isomap, MDS, Eigenmaps, and Spectral Clustering," in *Advances in Neural Information Processing Systems*. MIT Press, 2003, pp. 177–184.

[67] M. W. Trosset and C. E. Priebe, "The out-of-sample problem for classical multidimensional scaling," *Computational Statistics and Data Analysis*, vol. 52, no. 10, pp. 4635–4642, 2008.

[68] G. Szekely and M. Rizzo, "Brownian distance covariance," *Annals of Applied Statistics*, vol. 3, no. 4, pp. 1233–1303, 2009.

[69] N. Mekuz and J. K. Tsotsos, "Parameterless Isomap with adaptive neighborhood selection," in *Proceedings of the 28th DAGM Symposium*. Springer, 2006, pp. 364–373.

[70] T. P. Minka, "Automatic choice of dimensionality for PCA," *NIPS 13*, pp. 598–604, 2001a.

[71] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[72] J. Plaat, "FastMap, MetricMap, and landmark MDS are all Nystrom algorithms," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005, pp. 261–268.

[73] L. van der Maaten, E. Postma, and H. van den Herik, "Dimensionality reduction:

A comparative review," in *Tilburg University Technical Report, TiCC-TR 2009-005*, 2009.

[74] Y. Goldberg and Y. Ritov, "Manifold learning: the price of normalization," *Journal of Machine learning research*, vol. 9, pp. 1909–1939, 2008.

Vita

Cencheng Shen received his BSc degree in Quantitative Finance from National University of Singapore in 2010. He was then enrolled in the Ph.D. program at Johns Hopkins University, Department of Applied Mathematics and Statistics. His research mainly focuses on statistics and data analysis under the supervision of Professor Carey Priebe.